

5 **SEQUENCE-DIRECTED DNA-BINDING MOLECULES COMPOSITONS AND METHODS**

10

15

Field of the Invention

The present invention relates to methods, systems, and kits useful for the identification of molecules that specifically bind to defined nucleic acid sequences. Also described are methods for designing molecules having the ability to bind defined nucleic acid sequences and compositions thereof.

20

Table of Contents:

References

Page No.:

Background of the Invention

Summary of the Invention

Brief Description of the Figures

Detailed Description of the Invention

I. Definitions

II. The Assay

A. General Considerations

B. Choosing and Testing an Appropriate DNA-Binding Protein

1. Criteria for Choosing an Appropriate DNA-Binding Protein

2. Testing DNA:Protein Interactions for Use in the Assay

a.) Other DNA: Protein Interactions Useful in the Method of the Present Invention

b.) The Use of UL9 Proteins in the Practice of the Present Invention

C. Preparation of Full Length UL9 and UL9-COOH Polypeptides

D. Functional Binding and Rate of Dissociation

E. *oriS* Flanking Sequence Variation

F. Small Molecules as Sequence-Specific Competitive Inhibitors

G. Theoretical Considerations on the Concentration of Assay Components

H. The Use of the Assay under Conditions of Very High Protein Concentration

III. Amplification-Based Selection Technique to Determine the Sequence Preferences of DNA-Binding Molecules

A. Design of Test Oligonucleotides

B. Applying the Assay to the Mixed Pools of Test Oligonucleotides

C. Amplification

D. Sequencing of Amplified DNA

IV. Modifications of Test Oligonucleotides and other Useful DNA:Protein Interactions

V. Capture/Detection Systems

A. Capture of Unbound DNA

1. Modification of the Protein Recognition Sequence with Biotin

55

60

65

	2.	Capture of Site-Specific Biotinylated Oligonucleotides	67
	B.	Capture of DNA: Protein Complexes	69
	C.	Detection Systems	69
5	1.	Radioactive Labeling	70
	2.	Chemiluminescent Detection	71
	D.	Alternative Methods for Detecting Molecules that Increase the Affinity of the DNA-Binding Protein for its Cognate Site	72
	VI.	Utility	73
10	A.	The Usefulness of Sequence-Specific DNA-Binding Molecules	73
	B.	General Applications of the Assay	75
	1.	Mass-Screening of Libraries for the Presence of Sequence-Specific DNA-Binding Molecules	76
	2.	Directed Screening	78
15	3.	Molecules Derived from Known DNA-binding Molecules	79
	4.	Secondary Assays	80
	a.)	Confirmatory Studies	80
	b.)	Secondary Studies to Elucidate Binding Characteristics	86
20	c.)	Restriction Endonucleases as Indicator Proteins in the Assay	88
	5.	Generation of Binding Data and Refinement of Molecular Modeling Systems	90
25	6.	The Design of New DNA-Binding Heteropolymers Comprised of Subunits Directed to Different DNA Sequences	91
	C.	Sequences Targeted by the Assay	92
	1.	Medically Significant Target Sequences	94
30	a.)	Table I: Pathogens	95
	b.)	Table II: Non-infectious Diseases	108
	c.)	Table III: Human Genes with Promoter Regions that are Potential Targets for DNA-Binding Molecules	114
35	d.)	Table IV: Medically Significant DNA-Binding Sequences	128
	2.	Defined Sets of Test Sequences	132
40	3.	Theoretical Considerations in Choosing Biological Target Sites: Specificity and Toxicity	133
	4.	Further Considerations in Choosing Target Sites: Finding Eukaryotic Promoters	133
45	5.	Further Considerations in Choosing Alternative Small-Molecule Binding Sites	136
	6.	Further Considerations in Choosing Target Sites: Prokaryotes and Viruses	139
	D.	Using Test Matrices and Pattern Matching for the Analysis of Data	143
50	E.	Applications for the Determination of the Sequence Specificity of DNA-Binding Drugs	146
	1.	Multimerization of Sequence-Preferential or Sequence-Specific DNA-Binding Molecules Identified in the Assay	147
55	2.	Sequence-Specific DNA-Binding Molecules Identified in the Assay as Facilitators of Triplex Formation	152
	F.	Other Applications	155
		Materials and Methods	156
		Examples	156
		Sequence Listing	211
60		Claims	534
		Abstract	541

References

- Ambinder, R.F., et al., *J. Virol.* 65:1466-1478 (1991).
 Angel, P., et al., *Nature* 332:166 (1988).
 Ausubel, F. M., et al., Current Protocols in Molecular Biology, John Wiley and Sons, Inc., Media PA.
 Baguley, B.C., *Mol. Cell. Bioch.* 43:167-181 (1982).
 Banerji, S.S., et al., *Mol. Cell Biol.* 11:4074-4087 (1991).
 Beal, P.A., et al., *Science* 251:1360-1363 (1991).
 Becker, Y., et al., *Isr. J. Med. Sci.* 8:1225 (1972).
 Bialer, M., et al., *J. Med. Chem.* 23:1144 (1980).
 Bialer, M., et al., *J. Pharm. Sci.* 70:822 (1981).
 Birg, F., et al., *Nucl. Acids Res.* 18:2901-2908 (1990).
 Bohmann, D., et al., *Science* 238:1386 (1987).
 Bos, T.J., et al., *Cell* 52:705 (1988).
 Chaiet, L., et al., *Arch. Biochem. Biophys.* 106:1 (1964).
 Chaires, J.B., et al., *Biochemistry* 29:6145-6153 (1990).
 Chang, H.-K., et al., *Mol. Cell. Biol.* November:5189-5197 (1989).
 Chen, K-X., et al., *J. Biomol. Struct. Dyn.* 3:445-466 (1985).
 Chin, M.T., et al., *J. Virol.* 63:2967-2976 (1989).
 Comai, L., et al., *Cell* 68:965-976 (1992).
 Cooney, M., et al., *Science* 241:456-459 (1988).
 Courtois, G., et al., *Proc. Natl. Acad. Sci. USA* 85:7937-7941 (1988).
 Cullinane, C., et al., *FEBS Lett.* 293:195-198 (1991).
 Debart, F., et al., *J. Med. Chem.* 32:1074 (1989).
 Dervan, P.B., *Science* 232:464-471 (1986).
 Descheemaeker, K.A., et al., *J. Biol. Chem.* 267(21):15086 (1992).
 Edwards, C.A. et al., *J. Mol. Biol.* 180:73-90 (1984).
 Edwards, C.A., et al., in: Advances in Regulation of Cell Growth, Volume I: Regulation of Cell Growth and Activation, edited by Mond, J.J., et al., New York: Raven Press, p. 91-118 (1989).
 Elias, P., et al., *Proc. Natl. Acad. Sci. USA* 85:2959-2963 (1988).
 Fox, K.R., et al., *Biochim. Biophys. Acta* 840:383-392 (1985).
 Fox, K.R., et al., *Nucl. Acids Res.* 16:2489-2507 (1988).
 Fox, K.R., et al., *Nucl. Acids Res.* 18:1957-1963 (1990).
 Fox, K.R., et al., *Biochem J.* 269:217-221.
 Fried, M.G., et al., *Nuc. Acid. Res.* 9:6505 (1981).
 Galas, D., et al., *Nuc. Acid Res.* 5:3157-3170 (1981).
 Garner, M.M., et al., *Nuc. Acid. Res.* 9:3047 (1981).
 Gaugain, B., et al., *Biochemistry* 17:5071 (1978).
 Gessner, R.V., et al., *Biochemistry* 24:237-240 (1985).
 Gilbert, D.F., et al., *Proc. Natl. Acad. Sci. USA* 86:3006 (1988).
 Gilman, A. G., et al., eds., The Pharmacological Basis of Therapeutics, Eighth Edition, Pergamon Press (1990).
 Goldin, A.L., et al., *J. Virol.* 38:5-58 (1981).
 Goodisman, J., et al., *Biochemistry* 31:1046-1058 (1992).
 Green, N.M., *Adv. Protein Chem.* 29:85 (1975).
 Greenblatt, J., *Cell* 66:1067-1070 (1991).
 Greene, W.C., *Annu. Rev. Immunol.* 8:453-475 (1990).
 Griffen, J.H., et al., *J. Am. Chem. Soc.* (1992).
 Griffin, L.C., et al., *Science* 245:967-971 (1989).
 Gross, D.S., et al., *Annu. Rev. Biochem.* 57:159-197 (1988).
 Gurskii, G.V., et al., *Mol. Biol.* 19:177 (1985).
 Harlow, E., et al., Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press (1988).
 Harshman, K.D., et al., *Cell* 53:321 (1988).
 Hausheer, F.H., et al., *Anti-Cancer Drug Design* 5:159-167 (1990).
 Hawley, R.C., et al., *Proc. Natl. Acad. Sci. USA* 86:1105-1109 (1989).
 Helene, C., et al., *Biochim. Biophys. Acta* 1049:99-125 (1990).
 Helene, C., et al., *Genome* 31:413-420 (1989).
 Hoogsteen, *Acta Cryst.* 12 822 (1959).
 Innis et al., eds. PCR Protocols, a Guide to Methods and Applications, Academic Press, Inc. (1991).
 Jain, S.C., et al., *J. Mol. Biol.* 68:1-20 (1972).
 Jeppesen, C., et al., *Eur. J. Biochem.* 182:437-444 (1989).

- Kadonaga, J.T., *PNAS* 83:5889-5893 (1986).
- Kissinger, K., et al., *Biochemistry* 26:5590 (1987).
- Kitadai, Y., et al., *Biochem. Biophys. Res. Commun.* 189(3):1342 (1992).
- Koff, A., et al., *J. Virol.* 62:4096-4103 (1988).
- Kotler, M., et al., *FEBS Lett.* 21:222 (1972).
- Krowicki, K., et al., *J. Org. Chem.* 52:3493 (1987).
- Kuhlmann, K.F., et al., *Nucl. Acids Res.* 5:2629 (1978).
- Laugaa, P., et al., *Biochemistry* 23:1336 (1985).
- Le Pecq, J.B., et al., *Proc. Natl. Acad. Sci. U.S.A.* 72:2915-2919 (1975).
- Lee, D.K., et al., *Cell* 67:1241-1250 (1991).
- Lown, J.W., et al., *J. Org. Chem.* 50:3774 (1985).
- Lown, J.W., et al., *J. Med. Chem.* 29:1210 (1986).
- Luck, G., et al., *Nucl. Acids Res.* 1:503 (1974).
- Luckow, V.A., et al., *Virology* 170:31 (1989).
- Maher III, L.J., et al., *Science* 245:725-730 (1989).
- Maher, L.J., et al., *Biochemistry* 31(1):70-81 (1992).
- Maniatis, T., et al., *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory (1982).
- Maxam, A.M., et al., *Meth. Enzymol.*, 65:499 (1980).
- McGeoch, D.J., et al., *J. Virol.* 62:444-453 (1988).
- Meijer, I., et al., *Cell-Immunol.* 145(1):56 (1992).
- Miller, et al., U.S. Patent No. 4,757,055, issued 19 July 1988.
- Montenay-Garestier, T., et al., *CIBA Found. Symp.* 158:147-157.
- Mullis, K.B., U.S. Patent No. 4,683,202, issued 28 July 1987.
- Mullis, K.B., et al., U.S. Patent No. 4,683,195, issued 28 July 1987.
- Nakamura, S., et al., *J. Antibiot., Ser. A.* 17:220 (1964).
- Neuberg, M., et al., *Oncogene* 6(8):1325 (1991).
- Olivo, P.D., et al., *Proc. Natl. Acad. Sci. USA* 85:5414-5418 (1988).
- Olivo, P.D., et al., *J. Virology* 3:196-204 (1989).
- Pelaprat, D., et al., *J. Med. Chem.* 23:1336-1343 (1980).
- Perouault, L., et al., *Nature* 344:358-360 (1990).
- Phillips, D.R., *Anti Cancer Drug Design* 5:21-29 (1990).
- Phillips, et al., *Biochemistry* 29:4812-4819 (1990).
- Pitha, *Biochem. Biophys. Acta* 204:39 (1970a).
- Pitha, *Biopolymers* 9:965 (1970b).
- Portugal, J., et al., *FEBS Lett.* 225:195-200 (1987).
- Quigley, G.J., et al., *Science* 232:1255-1258 (1986).
- Raney, A.K., et al., *J. Virol.* 66(12):6912 (1992).
- Reisman, D., et al., *Mol. Cell. Biol.* 5:1822-1832 (1985).
- Remers, W.A., *Antineoplastic Agents*, New York: John Wiley and Sons, Inc., 1992.
- Rice, J.A., et al., *Proc. Natl. Acad. Sci. USA* 85:4158-4161 (1988).
- Ryder, K., et al., *Proc. Natl. Acad. Sci. USA* 85:1487 (1988).
- Salas, X., et al., *FEBS Lett.* 292:223-228 (1991).
- Sambrook, J., et al., *In Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Vol. 2 (1989).
- Sanger, F., et al., *Proc Nat Acad Sci, USA*, 74:5363 (1977).
- Schmidt, A., et al., *J. Virol.* 64:4037-4041 (1990).
- Schultz, P.G., et al., *Proc. Natl. Acad. Sci. USA* 80:6834-6837 (1983).
- Schuhmann, E., et al., *Allg. Microbiol.* 14:321 (1974).
- Shaw, J.P., et al., *Science* 241:202 (1988).
- Sherman, S.E., et al., *Chem. Rev.* 87:1153 (1987).
- Siebenlist, U., et al., *Proc. Natl. Acad. Sci. USA* 77:122-126 (1980).
- Skorobogaty, A., et al., *Anti-Cancer Drug Design* 3:41-56 (1988).
- Smith, D.B., et al., *Gene* 67:31 (1988).
- Sobell, H.M., et al., *J. Mol. Biol.* 68:21-34 (1972).
- Sobell, H.M., *Prof. Nucl. Acid. Res. Mol. Biol.* 13:153-190 (1973).
- Stow, N.D., et al., *Virology* 130:427-438 (1983).
- Stow, N.D., et al., *J. Gen. Virol.* 67:1613-1623 (1986).
- Strobel, S.A., et al., *Science* 249:73-75 (1990).
- Summers, M.D., et al., *A Manual of Methods for Baculovirus Vectors and Insect Cell Culture Procedures*, Texas Agricultural Experimental Station Bulletin, No. 1555 (1987).
- Summerton, J., et al., PCT International Application, Publication No. WO

86/05518, Published 25 September 1986.

Summerton, J., et al., U.S. Patent No. 5,034,506, issued 23 July 1991.

Thompson, C.B., et al., *Molecular and Cell Biology* 12(3):1043 (1992).

Thrum, H., et al., *Antimicrobial and Antineoplastic Chemotherapy*,
5 Prague: Czech. Med. Press, pp. 819-822 (1972).

Tullius, T.D., *Ann. Rev. Biophys. Biochem.* 18:213-237 (1989).

Wang, A.H.-J., et al., *Science* 225:1115-1121 (1984).

Wartel, R.M., et al., *J. Biol. Chem.* 15:285-318 (1975).

Weir, H.M., et al., *Nucl. Acids Res.* 17:1409-1425 (1989).

10 Werner, G.H., et al., *Actual. Pharmaceut. Fr.* 21:133 (1963).

White, R.J., et al., *Biochemistry* 28:6259-6269 (1989).

Wirth, T., et al., *EMBO J.* 7(10):3109 (1988).

Woodbury, C.P., et al., *Biochemistry* 22(20):4730-4737 (1983).

Wu, C.A., et al., *J. Virol.* 62:435-443 (1988).

15 Young, S.L., et al., *Proc. Natl. Acad. Sci. U.S.A.* 88:10023-10026 (1991).

Zein, N., et al., *Science* 240:1198 (1988).

Zimmer, C., *Pros. Nucl. Acid Res. Mol. Biol.* 15:285-318 (1975).

Background of the Invention

20 Several classes of small molecules that interact with double-stranded DNA have been identified. Many of these small molecules have profound biological effects. For example, many aminoacridines and polycyclic hydrocarbons bind DNA and are mutagenic, teratogenic, or carcinogenic. Other small molecules that bind DNA include: biological metabolites, some of which have applications as
25 antibiotics and antitumor agents including actinomycin D, echinomycin, distamycin, and calicheamicin; planar dyes, such as ethidium and acridine orange; and molecules that contain heavy metals, such as cisplatin, a potent antitumor drug.

The sequence binding preferences of most known DNA binding molecules have not, to date, been identified. However, several small DNA-binding molecules have been shown to preferentially recognize specific nucleotide sequences, for example: echinomycin has been shown to preferentially bind the sequence [(A/T)CGT]/[ACG(A/T)] (Gilbert et al.); cisplatin has been shown to covalently cross-link a platinum molecule between the N7 atoms of two adjacent
30 deoxyguanosines (Sherman et al.); and calicheamicin has been shown to preferentially bind and cleave the sequence TCCT/AGGA (Zein et al.).

Many therapeutic DNA-binding molecules (such as distamycin) that were initially identified based on their therapeutic activity in a biological screen have been later determined to bind DNA. There are several examples in the literature referring to synthetic or naturally-occurring polymers of DNA-binding drugs. Netropsin, for example, is a naturally-occurring oligopeptide that binds to the minor groove of double-stranded DNA. Netropsin contains two 4-amino-1-methylpyrrole-2-carboxylate residues and belongs to a family of similar biological metabolites from *Streptomyces* spp. This family includes
35 distamycin, anthelvencin (both of which contain three N-methylpyrrole residues), noformycin, amidomycin (both of which contain one N-methylpyrrole residue) and kikumycin (which contains two N-methylpyrrole residues, like netropsin) (Debart, et al.). Synthetic molecules of this family have also been described, including the above-mentioned molecules (Lown, et al. 1985) well as dimeric derivatives (Griffin et al., Gurskii, et al.) and certain analogues (Bialer, et al. 1980, Bialer, et al. 1981, Krowicki, et al.).

Molecules in this family, particularly netropsin and distamycin, have been of interest because of their biological activity as antibacterial (Thrum et al., Schuhmann, et al.), antiparasitic (Nakamura et al.), and antiviral
40 drugs (Becker, et al., Lown, et al. 1986, Werner, et al.).

Among the synthetic analogs of netropsin and distamycin are oligopeptides that have been designed to have sequence preferences different from their parent molecules. Such oligopeptides include the "lexitropsin" series of analogues. The N-methylpyrrole groups of the netropsin series were
45 systematically replaced with N-methylimidazole residues, resulting in lexitropsins with increased and altered sequence specificities from the parent compounds (Kissinger, et al.). Further, a number of poly(N-methylpyrrolyl)-netropsin analogues have been designed and synthesized which extend the number of residues in the oligopeptides to increase the size of the binding site
50
55
60

(Dervan, 1986).

There are several different approaches that could be taken to look for small molecules that specifically inhibit the interaction of a given DNA-binding protein with its binding sequence (cognate site). One approach would be to test biological or chemical compounds for their ability to preferentially block the binding of one specific DNA:protein interaction but not others. Such an assay would depend on the development of at least two, preferably three, DNA:protein interaction systems in order to establish controls for distinguishing between general DNA-binding molecules (polycations like heparin or intercalating agents like ethidium) and DNA-binding molecules having sequence binding preferences that would affect protein/cognate binding site interactions in one system but not the other(s).

One illustration of how this system could be used is as follows. Each cognate site could be placed 5' to a reporter gene (such as genes encoding β -galactoside or luciferase) such that binding of the protein to the cognate site would enhance transcription of the reporter gene. The presence of a sequence-specific DNA-binding drug that blocked the DNA:protein interaction would decrease the enhancement of the reporter gene expression. Several DNA enhancers could be coupled to reporter genes, then each construct compared to one another in the presence or absence of small DNA-binding test molecules. In the case where multiple protein/cognate binding sites are used for screening, a competitive inhibitor that blocks one interaction but not the others could be identified by the lack of transcription of a reporter gene in a transfected cell line or in an *in vitro* assay. Only one such DNA-binding sequence, specific for the protein of interest, could be screened with each assay system.

This approach has a number of limitations including limited testing capability and the need to construct the appropriate reporter system for each different protein/cognate site of interest.

Another example of a system to detect sequence-specific DNA-binding molecules would involve cloning a DNA-binding protein of interest, expressing the protein in an expression system (e.g., bacterial, baculovirus, or mammalian expression systems), preparing a purified or partially purified sample of protein, then using the protein in an *in vitro* competition assay to detect molecules that blocked the DNA:protein interaction. These types of systems are analogous to many receptor:ligand or enzyme:substrate screening assays developed in the past, but have the same limitations as outlined above in that a new system must be developed for every different protein/cognate site combination of interest. The capacity for screening numerous different sequences is therefore limited.

Another example of a system designed to detect sequence-specific DNA-binding drugs would be the use of DNA footprinting procedures as described in the literature. These methods include DNase I or other nuclease footprinting (Chaires, et al.), hydroxy radical footprinting (Portugal, et al.), methidiumpropyl EDTA(iron) complex footprinting (Schultz, et al.), photofootprinting (Jeppesen, et al.), and bidirectional transcription footprinting (White, et al.). These procedures are likely to be accurate within the limits of their sequence testing capability but are seriously limited by (i) the number of different DNA sequences that can be used in one experiment (typically one test sequence that represents the binding site of the DNA-binding protein under study), and (ii) the difficulty of developing high throughput screening systems.

Summary of the Invention

In one aspect, the invention includes a method of constructing a DNA-binding agent capable of sequence-specific binding to a duplex DNA target region. The method includes identifying in the duplex DNA, a target region containing a series of at least two non-overlapping base-pair sequences of four base-pairs each, where the four base-pair sequences are adjacent, and each sequence is characterized by sequence-preferential binding to a duplex DNA-binding small molecule. The small molecules are coupled to form a DNA-binding agent capable of sequence-specific binding to said target region.

In one embodiment, the duplex-binding small molecules are identified as molecules capable of binding to a selected test sequence in a duplex DNA by first adding a molecule to be screened to a test system composed of (a) a DNA-

binding protein that is effective to bind to a screening sequence in a duplex DNA, with a binding affinity that is substantially independent of the test sequence adjacent the screening sequence, but that is sensitive to binding of molecules to such test sequence, when the test sequence is adjacent the screening sequence, and (b) a duplex DNA having said screening and test sequences adjacent one another, where the binding protein is present in an amount that saturates the screening sequence in the duplex DNA.

The test molecule is incubated in the test system for a period sufficient to permit binding of the molecule being tested to the test sequence in the duplex DNA. The degree of binding protein bound to the duplex DNA before adding the test molecule is compared with that after adding the molecule. The screening sequence may be from the HSV origin of replication, and the binding protein may be UL9. Exemplary screening sequences are identified as SEQ ID NO:601, SEQ ID NO:602, SEQ ID NO:615, and SEQ ID NO:641.

Specific examples of tetrameric basepair sequences include TTTC, TTTC, TTAC, TTAG, TTGC, TTGG, TTCC, TTCG, TATC, TATG, TAAC, TAAG, TAGC, TAGG, TACC, TAGC sequences. A specific example of a small molecule capable of binding to these sequences is distamycin.

In another aspect, the invention includes a method of blocking transcriptional activity from a duplex DNA template. The method includes identifying in the duplex DNA, a binding site for a transcription factor and, adjacent the binding site, a target region having a series of at least two non-overlapping tetrameric base-pair sequences, where the four (tetrameric) base-pair sequences are adjacent and each sequence is characterized by sequence-preferential binding to a duplex DNA-binding small molecule. The sequences are contacted with a binding agent composed of the small molecules coupled to form a DNA-binding agent capable of sequence-specific binding to said target region.

The target may be selected, for example, from DNA sequences adjacent a binding site for a eucaryotic transcription factor, such as transcription factor TFIID, or a procaryotic transcription factor, such as transcription sigma factor.

For mammalian transcription factors, the target region is typically chosen from non-conserved regions adjacent the transcription factor binding site. Target regions can be chosen so that the small molecule binding overlaps an adjacent transcription factor DNA binding sequence (e.g., for a TFIID binding site, by 1-3 nucleotide pairs). In this case, the specificity of DNA binding for the small molecule is essentially derived from the non-conserved sequences adjacent the transcription factor binding site, in order to reduce small molecule binding at the transcription factor binding site associated with other genes.

Also disclosed is a DNA-binding agent capable of binding with base-sequence specificity to a target region in duplex DNA, where the target region contains at least two adjacent four base-pair sequences. The agent includes at least two subunits, where each subunit is a small molecule which has a sequence-preferential binding affinity for a sequence of four base-pairs in the target region. The subunits are coupled to form a DNA-binding agent capable of sequence-specific binding to said target region.

In one general embodiment, the agent is designed for binding to a sequence in which the two tetrameric basepair sequences are separated (for example, by up to 20 basepairs, typically, 1 to 6 basepairs) and the small molecules in the agent are coupled to each other by a spacer molecule.

Also forming part of the invention is a method of constructing a binding agent capable of sequence-specific binding to a duplex DNA target region. The method includes identifying in the duplex DNA, a target region containing (i) a series of at least two adjacent non-overlapping base-pair sequences of four base-pairs each, where each four base-pair sequence is characterized by sequence-preferential binding to a duplex DNA-binding small molecule, and (ii) adjacent to (i) a DNA duplex region capable of forming a triplex with a third-strand oligonucleotide. The two small molecules are coupled to form a DNA-binding agent capable of sequence-specific binding to said target region, and the DNA-binding agent is attached to a third-strand oligonucleotide.

The binding of the DNA-binding agent to duplex DNA causes a shift from B form to A form DNA, allowing triplex binding between the third-strand polynucleotide and a portion of the target sequence.

Also disclosed is a triple-strand forming agent for use in practicing the method.

In still another aspect, the invention includes a method of ordering the sequence binding preferences a DNA-binding molecule. The method includes adding a molecule to be screened to a test system composed of (a) a DNA-binding protein that is effective to bind to a screening sequence in a duplex DNA with a binding affinity that is substantially independent of such test sequence adjacent the screening sequence, but that is sensitive to binding of molecules to such test sequence, and (b) a duplex DNA having said screening and test sequences adjacent one another, where the binding protein is present in an amount that saturates the screening sequence in the duplex DNA. The molecule in the test system is incubated for a period sufficient to permit binding of the molecule being tested to the test sequence in the duplex DNA, and the amount of binding protein bound to the duplex DNA before and after addition of the test molecule is compared. These steps are repeated using all test sequences of interest, and the sequences are then ordered on the basis of relative amounts of protein bound in the presence of the molecule for each test sequence.

The test sequences are selected, for example, from the group of 256 possible four base sequences composed of A, G, C and T. The DNA screening sequence is preferably from the HSV origin of replication, and the binding protein is preferably UL9.

The invention also includes, a method for altering the binding characteristics of a DNA-binding protein to a duplex DNA. In the method, a binding site for the DNA-binding protein is identified in the duplex DNA and a target region identified adjacent the binding site. A small molecule is selected that is characterized by sequence-preferential binding to the target region. Such molecules can be selected by the assay and methods of the present invention. When the small molecule is bound to the target region, the small molecule is typically adjacent to the binding site for the DNA-binding protein. Alternatively, the binding of the small molecule may overlapping the site for the DNA-binding protein by at least one nucleotide pair. In the case of such overlap, the specificity of DNA binding for the small molecule is essentially derived from non-conserved sequences adjacent the DNA-binding protein's binding site -- in order to reduce small molecule binding at similar DNA:protein binding sites at other locations. Finally, the duplex DNA is contacted with the small molecule at a concentration effective to alter binding of the DNA-binding protein to its binding site.

In this method, contacting the duplex DNA with a small molecule can either inhibit or enhance the binding of the DNA-binding protein to its binding site: depending on the small molecule that is selected. Exemplary DNA binding proteins include DNA replication factors and a variety of transcription factors.

One application of this method is to eucaryotic general transcription factors (e.g., TFIID), where the target region is typically selected from DNA sequences adjacent the binding site for the eucaryotic transcription factor (e.g., SEQ ID NO:1 to SEQ ID NO:600). In one embodiment, the DNA binding protein is a eucaryotic general transcription factor and the small molecule binds, in addition to the target region, 1 to three nucleotide pairs of the DNA-binding protein's binding site. In the case of TFIID, the small molecule typically binds to (i) the target region, and (ii) up to two nucleotides of the binding site for TFIID, where the nucleotides are contiguous to the target region.

Generally, the present invention provides a method of screening for molecules capable of binding to a selected test sequence in a duplex DNA. In the method of the present invention a test sequence of interest is selected. Such sequences can be selected, for example, from the group of sequences presented as SEQ ID NO:1 to SEQ ID NO:600. Alternatively, the test sequences can be sequences having randomly generated sequences or defined sets of sequences, such as, the group of 256 possible four base sequences composed of A, G, C and T.

A duplex DNA test oligonucleotide is constructed having a screening sequence adjacent a selected test sequence, where a DNA binding protein is effective to bind to the screening sequence with a binding affinity that is substantially independent of the adjacent test sequence. In such constructs

the DNA protein binding to the screening sequence is sensitive to binding of test molecules to the test sequence.

Molecules selected for testing/screening are added to a test system composed of (a) the DNA binding protein, and (b) the duplex DNA test oligonucleotide, which contains the screening and test sequences adjacent one another. Selected molecules are incubated in the test system for a period sufficient to permit binding of the molecule being tested to the test sequence in the duplex DNA. The amount of binding protein bound to the duplex DNA is compared before and after adding a test molecule. Comparison of the amount of binding protein bound to the duplex DNA before and after adding a test molecule can be accomplished, for example, using a gel band-shift assay or a filter-binding assay.

In the method of the present invention a number of DNA:protein interactions may be used for screening purposes. In one embodiment, the DNA screening sequence is from the HSV origin of replication and the binding protein is UL9. Exemplary HSV origin of replication screening sequences include SEQ ID NO:601, SEQ ID NO:602, SEQ ID NO:615, and SEQ ID NO:641.

Other DNA:protein interactions useful in the practice of the present invention include restriction endonucleases and their cognate DNA-binding sequences. These reactions are typically carried out in the absence of divalent cations.

In another embodiment, the invention includes a method of identifying test sequences in duplex DNA to which binding of a test molecule is most preferred. In this method a mixture of duplex DNA test oligonucleotides is constructed, where each oligonucleotide has a screening sequence adjacent a test sequence as described above. The test oligonucleotides of the mixture typically contain different test sequences.

A test molecule, to be screened, is added to a test reaction composed of (a) the DNA binding protein, and (b) the duplex DNA test oligonucleotide mixture. The molecule is incubated in the test reaction for a period sufficient to permit binding of the compound being tested to test sequences in the duplex DNA. Test oligonucleotides are separated from test oligonucleotides bound to binding protein.

The test oligonucleotides can be separated from test oligonucleotides bound to protein by, for example, passing the test reaction through a filter, where the filter is capable of capturing DNA:protein complexes but not DNA that is free of protein. One filter type useful in the practice of the present invention is the nitrocellulose filter.

The separated test oligonucleotides are then amplified. These amplified test oligonucleotides are then recycled through the screening steps of the assay in order to obtain a desired degree of selection. The amplified test oligonucleotides are isolated and sequenced.

Exemplary test sequences include sequences selected from the group of 256 possible four base sequences composed of A, G, C and T. Further examples of desirable test sequences include test sequences derived from the sequences presented as SEQ ID NO:1 to SEQ ID NO:600.

The amplification step in the method may be accomplished by polymerase chain reaction or other methods of amplification, including, cloning and subsequent *in vivo* amplification of the cloning vector containing the sequences of interest.

These and other objects and features of the invention will be more fully appreciated when the following detailed description of the invention is read in conjunction with the accompanying drawings.

Brief Description of the Figures

Figure 1A illustrates a DNA-binding protein binding to a screening sequence. Figures 1B and 1C illustrate how a DNA-binding protein may be displaced or hindered in binding by a small molecule by two different mechanisms: because of steric hindrance (1B) or because of conformational (allosteric) changes induced in the DNA by a small molecule (1C).

Figure 2 illustrates an assay for detecting inhibitory molecules based on their ability to preferentially hinder the binding of a DNA-binding protein to its binding site. Protein (O) is displaced from DNA (I) in the presence of inhibitor (X). Two alternative capture/detection systems are illustrated, the capture and detection of unbound DNA or the capture and detection of

DNA:protein complexes.

Figure 3 shows a DNA-binding protein that is able to protect a biotin moiety, covalently attached to an oligonucleotide sequence, from being recognized by streptavidin when a protein is bound to the DNA.

Figure 4 shows the incorporation of biotin and digoxigenin into a typical oligonucleotide molecule for use in the assay of the present invention. The oligonucleotide contains the binding sequence (i.e., the screening sequence) of the UL9 protein, which is underlined, and test sequences flanking the screening sequence. Figure 4 also shows the preparation of double-stranded oligonucleotides end-labeled with either digoxigenin or ³²P.

Figure 5 shows a series of sequences that have been tested in the assay of the present invention for the binding of sequence-specific small molecules.

Figure 6 outlines the clonings, into an expression vector, of a truncated form of the UL9 protein (UL9-COOH) which retains its sequence-specific DNA-binding ability.

Figure 7 shows the pVL1393 baculovirus vector containing the full length UL9 protein coding sequence.

Figure 8 is a photograph of a SDS-polyacrylamide gel showing (i) the purified UL9-COOH/glutathione-S-transferase fusion protein and (ii) the UL9-COOH polypeptide.

Figure 9 presents data demonstrating the effect on UL9-COOH binding of alterations in the test sequences that flank the UL9 screening sequence.

Figure 10A shows the effect of the addition of several concentrations of distamycin A to DNA:protein assay reactions utilizing different test sequences.

Figure 10B shows the effect of the addition of actinomycin D to DNA:protein assay reactions utilizing different test sequences. Figure 10C shows the effect of the addition of Doxorubicin to DNA:protein assay reactions utilizing different test sequences.

Figure 11A illustrates a DNA capture system of the present invention utilizing biotin and streptavidin coated magnetic beads. The presence of the DNA is detected using an alkaline-phosphatase substrate that yields a chemiluminescent product. Figure 11B shows a similar reaction using biotin coated agarose beads that are conjugated to streptavidin, that in turn is conjugated to the captured DNA.

Figure 12 demonstrates a test matrix based on DNA:protein-binding data.

Figure 13 lists the top strands (5'-3') of all the possible four base pair sequences that could be used as a defined set of ordered test sequences in the assay.

Figure 14A lists the top strands (5'-3') of all the possible four base pair sequences that have the same base composition as the sequence 5'-GATC-3'.

This is another example of a defined, ordered set of sequences that could be tested in the assay. Figure 14B presents the general sequence of a test oligonucleotide (SEQ ID NO:617), where XXXX is the test sequence and N = A,G,C, or T.

Figure 15 shows the results of 4 duplicate experiments in which the binding activity of distamycin was tested with all possible (256) four base pair sequences. The oligonucleotides are ranked from 1 to 256 (column 1, "rank") based on their average rank from the four experiments (column 13, "ave. rank"). (rank is shown in the first column of the chart).

Figure 16 shows the average ranks (Figure 15) plotted against the ideal ranks 1 to 256.

Figure 17 shows the average r% scores (Figure 15) plotted against the rank of 1 to 256.

Figure 18 shows the results of eight experiments with actinomycin D. The r% scores and rank are shown for each of the 256 oligonucleotides.

Figure 19 shows the average r% versus rank, by average rank (data from Figure 18).

Figure 20 shows the ideal and average ranks for each of the 256 oligonucleotides.

Figure 21 shows the results of a position analysis for actinomycin D preference.

Figure 22 presents the data for a dinucleotide analysis of actinomycin D binding preference.

Figure 23 graphically displays the results presented in Figure 22.

Figure 24 graphically displays the data presented in Figure 22, where the

data are combined in a combined bar chart so that the cumulative results for any dinucleotide pair are tabulated in a single bar.

Figure 25 shows the top strands of 16 possible duplex DNA target sites for binding bis-distamycins.

Figure 26 shows examples of bis-distamycin target sequences for bis-distamycins with internal flexible and/or variable length linkers targeted to sites comprised of two TTCC sequences, where N is any base.

Figures 27A to 27H show sample oligonucleotides for competition binding studies using the assay of the present invention.

Figure 28 shows the DNA sequences of the HIV pro-viral promoter region. Several transcription factor binding sites are marked.

Figures 29A to 29D illustrate sample test oligonucleotides for use in the polymerase chain reaction based selection technique of the present invention. In Figure 29A, X is the number of bases that comprise the test site.

Figure 30 illustrates a sample test oligonucleotide for use in the assay of the present invention, where the test oligonucleotide employs several different DNA:protein interaction systems.

Figure 31 illustrates the results of screening a selected test sequence with a single DNA:protein interaction system. In the figure, the test site is shown in bold, the potential binding site for the test molecule is underlined.

Figure 32 illustrates the results of screening the same selected test sequence as shown in Figure 31, but using a different single DNA:protein interaction system. In the figure, the test site is shown in bold, the potential binding site for the test molecule is underlined.

Detailed Description of the Invention

I. Definitions:

Adjacent is used to describe the distance relationship between two neighboring sites. Adjacent sites are 20 or less bp apart, and can be separated by any fewer number of bases including the situation where the sites are immediately abutting one another. "Flanking" is a synonym for adjacent.

Bound DNA, as used in this disclosure, refers to the DNA that is bound by the protein used in the assay (e.g., a test oligonucleotide containing the UL9 binding sequence bound to the UL9 protein).

Coding sequences or coding regions are DNA sequences that code for RNA transcripts, unless specified otherwise.

Dissociation is the process by which two molecules cease to interact: the process occurs at a fixed average rate under specific physical conditions.

Functional binding is the noncovalent association of a protein or small molecule to the DNA molecule. In one embodiment of the assay of the present invention the functional binding of the UL9 protein to a screening sequence (i.e., its cognate DNA binding site) has been evaluated using filter binding or gel band-shift experiments.

Half-life is herein defined as the time required for one-half of the associated complexes, e.g., DNA:protein complexes, to dissociate.

Heteropolymers are molecules comprised of at least two different subunits, each representing a different type or class of molecule. The covalent coupling of different subunits, such as, DNA-binding molecules or portions of DNA-binding molecules, results in the formation of a heteropolymer: for example, the coupling of a non-intercalating homopolymeric DNA-binding molecule, such as distamycin, to an intercalating drug, such as daunomycin. Likewise, the coupling of netropsin, which is essentially a molecular subunit of distamycin, to daunomycin would also be a heteropolymer. As a further example, the coupling of distamycin, netropsin, or daunomycin to a DNA-binding homopolymer, such as a triplex-forming oligonucleotide, would result in a heteropolymer.

Homopolymers are molecules that are comprised of a repeating subunit of the same type or class. Two examples of duplex DNA-binding homopolymers are as follows: (i) triplex-forming oligonucleotides or oligonucleotide analogs, which are composed of repeating subunits of nucleotides or nucleotide analogs, and (ii) oligopeptides, which are composed of repeating subunits linked by peptide bonds (e.g., distamycin, netropsin).

Sequence-preferential binding refers to DNA binding molecules that generally bind DNA but that show preference for binding to some DNA sequences over others. Sequence-preferential binding is typified by several of the small

molecules tested in the present disclosure, e.g., distamycin. Sequence-preferential and sequence-specific binding can be evaluated using a test matrix such as is presented in Figure 12. For a given DNA-binding molecule, there are a spectrum of differential affinities for different DNA sequences ranging from non-sequence-specific (no detectable preference) to sequence preferential to absolute sequence specificity (i.e., the recognition of only a single sequence among all possible sequences, as is the case with many restriction endonucleases).

Sequence-specific binding refers to DNA binding molecules which have a strong DNA sequence binding preference. For example, the following demonstrate typical sequence-specific DNA-binding: (i) multimers (heteropolymers and homopolymers) of the present invention (e.g., Section IV.E.1, Multimerization; Example 13), and (ii) restriction enzymes and the proteins listed in Table IV.

Screening sequence is the DNA sequence that defines the cognate binding site for the DNA binding protein: in the case of UL9, the screening sequence can, for example, be SEQ ID NO:601.

Small molecules are desirable as therapeutics for several reasons related to drug delivery, including the following: (i) they are commonly less than 10 K molecular weight; (ii) they are more likely to be permeable to cells; (iii) unlike peptides or oligonucleotides, they are less susceptible to degradation by many cellular mechanisms; and, (iv) they are not as apt to elicit an immune response. Many pharmaceutical companies have extensive libraries of chemical and/or biological mixtures, often fungal, bacterial, or algal extracts, that would be desirable to screen with the assay of the present invention. Small molecules may be either biological or synthetic organic compounds, or even inorganic compounds (i.e., cisplatin).

Test sequence is a DNA sequence adjacent the screening sequence. The assay of the present invention screens for molecules that, when bound to the test sequence, affect the interaction of the DNA-binding protein with its cognate binding site (i.e., the screening sequence). Test sequences can be placed adjacent either or both ends of the screening sequence. Typically, binding of molecules to the test sequence interferes with the binding of the DNA-binding protein to the screening sequence. However, some molecules binding to these sequences may have the reverse effect, causing an increased binding affinity of the DNA-binding protein to the screening sequence. Some molecules, even while binding in a sequence specific or sequence preferential manner, might have no effect in the assay. These molecules would not be detected in the assay.

Unbound DNA, as used in this disclosure, refers to the DNA that is not bound by the protein used in the assay (i.e., in the examples of this disclosure, the UL9 protein).

II. The Assay.

One feature of the present invention is that it provides an assay to identify small molecules that will bind in a sequence-specific manner to medically significant DNA target sites. The assay facilitates the development of a new field of pharmaceuticals that operates by interfering with specific DNA functions, such as crucial DNA:protein interactions. A sensitive, well-controlled assay has been developed (i) to detect DNA-binding molecules and (ii) to determine their sequence-specificity and affinity. The assay can be used to screen large biological and chemical libraries. For example, the assay will be used to detect sequence-specific DNA-binding molecules in fermentation broths or extracts from various microorganisms.

Furthermore, another application for the assay is to determine the sequence specificity and relative affinities of known DNA-binding drugs (and other DNA-binding molecules) for different DNA sequences. Such drugs, which are currently used primarily as antibiotics or anticancer drugs, may have previously unidentified activities that make them strong candidates for therapeutics or therapeutic precursors in entirely different areas of medicine.

The use of the assay to determine the sequence-binding preference of these known DNA-binding molecules enables the rational design of novel DNA-binding molecules with enhanced sequence-binding preference. The methods for designing and testing these novel DNA-binding molecules is described below.

The screening assay of the present invention is basically a competition assay that is designed to test the ability of a test molecule to compete with a DNA-binding protein for binding to a short, synthetic, double-stranded oligodeoxynucleotide that contains the recognition sequence for the DNA-binding protein flanked on either or both sides by a variable test site. The variable test site may contain any DNA sequence that provides a reasonable recognition sequence for a DNA-binding test molecule. Molecules that bind to the test site alter the binding characteristics of the protein in a manner that can be readily detected. The extent to which such molecules are able to alter the binding characteristics of the protein is likely to be directly proportional to the affinity of the test molecule for the DNA test site. The relative affinity of a given molecule for different oligonucleotide sequences at the test site (i.e., test sequences) can be established by examining the molecule's effect on the DNA:protein interaction using each of the test sequences.

The assay can be used to test specific target sequences and to identify novel DNA-binding molecules. Also, the assay provides a means for the determination of the high affinity DNA binding sites for a given DNA-binding molecule, thus facilitating the identification of specific target sequences.

A. General Considerations.

The assay of the present invention has been designed for detecting test molecules or compounds that affect the rate of transfer of a specific DNA molecule from one protein molecule to another identical protein in solution.

A mixture of DNA and protein is prepared in solution. The concentration of protein is in excess to the concentration of the DNA so that virtually all of the DNA is found in DNA:protein complexes. The DNA is a double-stranded oligonucleotide that contains the recognition sequence for a specific DNA-binding protein (i.e., the screening sequence). The protein used in the assay contains a DNA-binding domain that is specific for binding to the sequence within the oligonucleotide. The physical conditions of the solution (e.g., pH, salt concentration, temperature) are adjusted such that the half-life of the complex is amenable to performing the assay (optimally a half-life of 5-120 minutes), preferably in a range that is close to normal physiological conditions.

As one DNA:protein complex dissociates, the released DNA rapidly reforms a complex with another protein in solution. Since the protein is in excess to the DNA, dissociations of one complex always result in the rapid reassociation of the DNA into another DNA:protein complex. At equilibrium, very few DNA molecules will be unbound. If the unbound DNA is the component of the system that is measured, the minimum background of the assay is the amount of unbound DNA observed during any given measurable time period. If the capture/detection system used for capturing the unbound DNA is irreversible, the brevity of the observation period (the length of time used to capture the unbound DNA) and the sensitivity of the detection system define the lower limits of background DNA.

Figure 1 illustrates how (i) such a protein can be displaced from its cognate binding site, (ii) a protein can be prevented from binding its cognate binding site, and (iii) how the kinetics of the DNA:protein interaction can be altered. In each case, the binding site for the test molecule is located at a site flanking the recognition sequence for the DNA-binding protein (Figure 1A).

One mechanism is steric hinderance of protein binding by a small molecule (competitive inhibition; Figure 1B). Alternatively, a molecule may interfere with a DNA:protein binding interaction by inducing a conformational change in the DNA (allosteric interference, noncompetitive inhibition; Figure 1C). In either event, if a test molecule that binds the oligonucleotide hinders binding of the protein, even transiently, the rate of transfer of DNA from one protein to another will be decreased. This will result in a net increase in the amount of unbound DNA and a net decrease in the amount of protein-bound DNA. In other words, an increase in the amount of unbound DNA or a decrease in the amount of bound DNA indicates the presence of an inhibitor, regardless of the mechanism of inhibition (competitive or noncompetitive).

Alternatively, molecules may be isolated that, when bound to the DNA, cause an increased affinity of the DNA-binding protein for its cognate binding site. In this case, the assay control samples (no drug added) are adjusted to less than 100% DNA:protein complex so that the increase in binding can be detected. The amount of unbound DNA (observed during a given measurable time

period after the addition of the molecule) will decrease and the amount of bound DNA will increase in the reaction mixture as detected by the capture/detection system described in Section II.

5 B. Choosing and Testing an Appropriate DNA-Binding Protein.

Experiments performed in support of the present invention have defined an approach for identifying molecules having sequence-preferential DNA-binding. In this approach small molecules binding to sequences adjacent the cognate binding sequence can inhibit the protein/cognate DNA interaction. This assay
10 has been designed to use a single DNA:protein interaction to screen for sequence-specific or sequence-preferential DNA-binding molecules that recognize virtually any sequence.

While DNA-binding recognition sites are usually quite small (4-17 bp), the sequence that is protected by the binding protein is larger (usually 5 bp or more on either side of the recognition sequence -- as detected by DNAase I protection (Galas, et al.) or methylation interference (Siebenlist, et al.).

Experiments performed in support of the present invention demonstrated that a single protein and its cognate DNA-binding sequence can be used to assay virtually any DNA sequence by placing a sequence of interest adjacent to the cognate site: a small molecule bound to the adjacent site can be detected by alterations in the binding characteristics of the protein to its cognate site.

Such alterations might occur by either steric hindrance (which would cause the dissociation of the protein) or induced conformational changes in the recognition sequence for the protein (which may cause either enhanced binding or, more likely, decreased binding of the protein to its cognate site).
25

1. Criteria for Choosing an Appropriate DNA-Binding Protein.

There are several considerations involved in choosing DNA:protein complexes that can be employed in the assay of the present invention including:

30 a.) The half-life of the DNA:protein complex should be short enough to accomplish the assay in a reasonable amount of time. The interactions of some proteins with their cognate binding sites in DNA can be measured in days not minutes: such tightly bound complexes would inconveniently lengthen the period of time it takes to perform the assay.

35 b.) The half-life of the complex should be long enough to allow the measurement of unbound DNA in a reasonable amount of time. For example, the level of free DNA is dictated by the ratio between the time needed to measure free DNA and the amount of free DNA that occurs naturally due to the dissociation of the complex during the measurement time period.

40 In view of the above two considerations, practical useful DNA:protein half-lives fall in the range of approximately two minutes to several days: shorter half-lives may be accommodated by faster equipment and longer half-lives may be accommodated by destabilizing the binding conditions for the assay.

45 c.) A further consideration is that the kinetic interactions of the DNA:protein complex is relatively insensitive to the nucleotide sequences flanking the recognition sequence. The affinity of DNA-binding proteins may be affected by differences in the sequences adjacent to the recognition sequence. If the half-life of the complex is affected by the flanking sequence, the analysis of comparative binding data between different flanking
50 oligonucleotide sequences becomes difficult but is not impossible.

2) Testing DNA:Protein Interactions for Use in the Assay.

55 a.) Other DNA:Protein Interactions Useful in the Method of the Present Invention. There are many known DNA:protein interactions that may be useful in the practice of the present invention, including (i) the DNA protein interactions listed in Table IV, (ii) bacterial, yeast, and phage systems such as lambda O_L - O_R /cro, and (iii) modified restriction enzyme systems (e.g., protein binding in the absence of divalent cations, see Section IV). Any
60 protein that binds to a specific recognition sequence may be useful in the present invention. One constraining factor is the effect of the immediately adjacent sequences (the test sequences) on the affinity of the protein for its recognition sequence. DNA:protein interactions in which there is little or no effect of the test sequences on the affinity of the protein for its cognate
65 site are preferable for use in the described assay; however, DNA:protein

interactions that exhibit test-sequence-dependent differential binding may still be useful if algorithms that compensate for the differential affinity are applied to the analysis of data. In general, the effect of flanking sequence composition on the binding of the protein is likely to be correlated to the length of the recognition sequence for the DNA-binding protein. That is, the kinetics of binding for proteins with shorter recognition sequences are more likely to suffer from flanking sequence effects, while the kinetics of binding for proteins with longer recognition sequences are more likely to not be affected by flanking sequence composition. The present disclosure provides methods and guidance for testing the usefulness of such DNA:protein interactions, in the screening assay.

b.) The Use of UL9 Proteins in the Practice of the Present Invention.

Experiments performed in support of the present invention have identified a DNA:protein interaction that is particularly useful for the above described assay: the Herpes Simplex Virus (HSV) UL9 protein that binds the HSV origin of replication (*oriS*). The UL9 protein has fairly stringent sequence specificity.

There appear to be three binding sites for UL9 in *oriS*, SEQ ID NO:601, SEQ ID NO:602 and SEQ ID NO:615 (Elias, et al.; Stow, et al.). One sequence (SEQ ID NO:601) binds with at least 10-fold higher affinity than the second sequence (SEQ ID NO:602): the embodiments described below use the higher affinity binding site (SEQ ID NO:601). Another useful UL9-binding site, *alibi* a lower affinity binding site, SEQ ID NO:641, has also been identified.

DNA:protein association reactions are performed in solution. The DNA:protein complexes can be separated from free DNA by any of several methods. One particularly useful method for the initial study of DNA:protein interactions has been visualization of binding results using band shift gels (Example 3A). In this method DNA:protein binding reactions are applied to polyacrylamide/TBE gels and the labelled complexes and free labeled DNA are separated electrophoretically. These gels are fixed, dried, and exposed to X-ray film. The resulting autoradiograms are examined for the amount of free probe that is migrating separately from the DNA:protein complex. These assays include (i) a lane containing only free labeled probe, and (ii) a lane where the sample is labeled probe in the presence of a large excess of binding protein. The band shift assays allow visualization of the ratios between DNA:protein complexes and free probe. However, they are less accurate than filter binding assays for rate-determining experiments due to the lag time between loading the gel and electrophoretic separation of the components.

The filter binding method is particularly useful in determining the half-life for oligonucleotide:protein complexes (Example 3B). In the filter binding assay, DNA:protein complexes are retained on a filter while free DNA passes through the filter. This assay method is more accurate for half-life determinations because the separation of DNA:protein complexes from free probe is very rapid. The disadvantage of filter binding is that the nature of the DNA:protein complex cannot be directly visualized. So if, for example, the competing molecule was also a protein competing for the binding of a site on the DNA molecule, filter binding assays cannot differentiate between the binding of the two proteins nor yield information about whether one or both proteins are binding.

c. Preparation of Full Length UL9 and UL9-COOH Polypeptides.

UL9 protein has been prepared by a number of recombinant techniques (Example 2). The full length UL9 protein has been prepared from baculovirus infected insect cultures (Example 3A, B, and C). Further, a portion of the UL9 protein that contains the DNA-binding domain (UL9-COOH) has been cloned into a bacterial expression vector and produced by bacterial cells (Example 3D and E).

The DNA-binding domain of UL9 is contained within the C-terminal 317 amino acids of the protein (Weir, et al.). The UL9-COOH polypeptide was inserted into the expression vector in-frame with the glutathione-S-transferase (*gst*) protein. The *gst*/UL9 fusion protein was purified using affinity chromatography (Example 3E). The vector also contained a thrombin cleavage site at the junction of the two polypeptides. Therefore, once the fusion protein was isolated (Figure 8, lane 2) it was treated with thrombin, cleaving the UL9-COOH/*gst* fusion protein from the *gst* polypeptide (Figure 8, lane 3). The UL9-COOH-*gst* fusion polypeptide was obtained at a protein purity of greater than

95% as determined using Coomassie staining.

Other hybrid proteins can be utilized to prepare DNA-binding proteins of interest. For example, fusing a DNA-binding protein coding sequence in-frame with a sequence encoding the thrombin site and also in-frame with the β -galactoside coding sequence. Such hybrid proteins can be isolated by affinity or immunoaffinity columns (Maniatis, et al.; Pierce, Rockford IL). Further, DNA-binding proteins can be isolated by affinity chromatography based on their ability to interact with their cognate DNA binding site. For example, the UL9 DNA-binding site (SEQ ID NO:601) can be covalently linked to a solid support (e.g., CnBr-activated Sepharose 4B beads, Pharmacia, Piscataway NJ), extracts passed over the support, the support washed, and the DNA-binding then isolated from the support with a salt gradient (Kadonaga). Alternatively, other expression systems in bacteria, yeast, insect cells or mammalian cells can be used to express adequate levels of a DNA-binding protein for use in this assay.

The results presented below in regard to the DNA-binding ability of the truncated UL9 protein suggest that full length DNA-binding proteins are not required for the DNA:protein assay of the present invention: only a portion of the protein containing the cognate site recognition function may be required. The portion of a DNA-binding protein required for DNA-binding can be evaluated using a functional binding assay (Example 4A). The rate of dissociation can be evaluated (Example 4B) and compared to that of the full length DNA-binding protein. However, any DNA-binding peptide, truncated or full length, may be used in the assay if it meets the criteria outlined in Section II.B.1, "Criteria for choosing an appropriate DNA-binding protein". This remains true whether or not the truncated form of the DNA-binding protein has the same affinity as the full length DNA-binding protein.

d. Functional Binding and Rate of Dissociation.

The full length UL9 and purified UL9-COOH proteins were tested for functional activity in "band shift" assays (see Example 4A). The buffer conditions were optimized for DNA:protein-binding (Example 4C) using the UL9-COOH polypeptide. These DNA-binding conditions also worked well for the full-length UL9 protein. Radiolabeled oligonucleotides (SEQ ID NO:614) that contained the 11 bp UL9 DNA-binding recognition sequence (SEQ ID NO:601) were mixed with each UL9 protein in appropriate binding buffer. The reactions were incubated at room temperature for 10 minutes (binding occurs in less than 2 minutes) and the products were separated electrophoretically on non-denaturing polyacrylamide gels (Example 4A).

The degree of DNA:protein-binding could be determined from the ratio of labeled probe present in DNA:protein complexes versus that present as free probe. This ratio was typically determined by optical scanning of autoradiograms and comparison of band intensities. Other standard methods may be used as well for this determination, such as scintillation counting of excised bands. The UL9-COOH polypeptide and the full length UL9 polypeptide, in their respective buffer conditions, bound the target oligonucleotide equally well.

The rate of dissociation was determined using competition assays. An excess of unlabelled oligonucleotide that contained the UL9 binding site was added to each reaction. This unlabelled oligonucleotide acts as a specific inhibitor, capturing the UL9 protein as it dissociates from the labelled oligonucleotide (Example 4B). The dissociation rate, as determined by a band-shift assay, for both full length UL9 and UL9-COOH was approximately 4 hours at 4°C or approximately 10 minutes at room temperature. Neither non-specific oligonucleotides (a 10,000-fold excess) nor sheared herring sperm DNA (a 100,000-fold excess) competed for binding with the oligonucleotide containing the UL9 binding site.

e. oriS Flanking Sequence Variation.

As mentioned above, one feature of a DNA:protein-binding system to be used in the assay of the present invention is that the DNA:protein interaction is not affected by the nucleotide sequence of the regions adjacent the DNA-binding site. The sensitivity of any DNA:protein-binding reaction to the composition of the flanking sequences can be evaluated by the functional binding assay and dissociation assay described above.

To test the effect of flanking sequence variation on UL9 binding to the

oriS SEQ ID NO:601 sequences oligonucleotides were constructed with 20-30 different sequences (i.e., the test sequences) flanking the 5' and 3' sides of the UL9 binding site. Further, oligonucleotides were constructed with point mutations at several positions within the UL9 binding site. Most point mutations within the binding site destroyed recognition. Several changes did not destroy recognition and these include variations at sites that differ between the UL9 binding sites (SEQ ID NO:601, SEQ ID NO:602, SEQ ID NO:615 and SEQ ID NO:641): the second UL9 binding site (SEQ ID NO:602) shows a ten-fold decrease in UL9:DNA binding affinity (Elias, et al.) relative to the first (SEQ ID NO:601). On the other hand, sequence variation at the test site (also called the test sequence), adjacent to the screening site (Figure 5, Example 5), had virtually no effect on binding or the rate of dissociation.

The results demonstrating that the nucleotide sequence in the test site, which flanks the screening site, has no effect on the kinetics of UL9 binding in any of the oligonucleotides tested is a striking result. This allows the direct comparison of the effect of a DNA-binding molecule on test oligonucleotides that contain different test sequences. Since the only difference between test oligonucleotides is the difference in nucleotide sequence at the test site(s), and since the nucleotide sequence at the test site has no effect on UL9 binding, any differential effect observed between the two test oligonucleotides in response to a DNA-binding molecule must be due solely to the differential interaction of the DNA-binding molecule with the test sequence(s). In this manner, the insensitivity of UL9 to the test sequences flanking the UL9 binding site greatly facilitates the interpretation of results. Each test oligonucleotide acts as a control sample for all other test oligonucleotides. This is particularly true when ordered sets of test sequences are tested (e.g., testing all 256 four base pair sequences (Figure 13) for binding to a single drug).

Taken together the above experiments support that the UL9-COOH polypeptide binds the SEQ ID NO:601 sequence with (i) appropriate strength, (ii) an acceptable dissociation time, and (iii) indifference to the nucleotide sequences flanking the screening site. These features suggested that the UL9/oriS system could provide a versatile assay for detection of small molecule/DNA-binding involving any number of specific nucleotide sequences.

The above-described experiment can be used to screen other DNA:protein interactions to determine their usefulness in the present assay.

f. Small Molecules as Sequence-Specific Competitive Inhibitors.

To test the utility of the present assay system several small molecules that have sequence-binding preferences (i.e., a preference for AT-rich versus GC-rich sequences) have been tested.

Distamycin A binds relatively weakly to DNA ($K_A = 2 \times 10^5 \text{ M}^{-1}$) with a preference for non-alternating AT-rich sequences (Jain, et al.; Sobell; Sobell, et al.). Actinomycin D binds DNA more strongly ($K_A = 7.6 \times 10^7 \text{ M}^{-1}$) than Distamycin A and has been reported to have a relatively strong preference for the dinucleotide sequence dGdC (Luck, et al.; Zimmer; Wartel). Each of these molecules poses a stringent test for the assay. Distamycin A tests the sensitivity of the assay because of its relatively weak binding. Actinomycin D challenges the ability to utilize flanking sequences since the UL9 recognition sequence contains a dGdC dinucleotide: therefore, it might be anticipated that all of the oligonucleotides, regardless of the test sequence flanking the assay site, might be equally affected by actinomycin D.

In addition, Doxorubicin, a known anti-cancer agent that binds DNA in a sequence-preferential manner (Chen, K-X, et al.), has been tested for preferential DNA sequence binding using the assay of the present invention.

Actinomycin D, Distamycin A, and Doxorubicin have been tested for their ability to preferentially inhibit the binding of UL9 to oligonucleotides containing different sequences flanking the UL9 binding site (Example 6, Figure 5). Furthermore, distamycin A and actinomycin D have been screened against all possible 256 4 bp DNA sequences. Binding assays were performed as described in Example 5. These studies were completed under conditions in which UL9 is in excess of the DNA (i.e., most of the DNA is in DNA:protein complexes).

In the preliminary studies, distamycin A was tested with 5 different test sequences flanking the UL9 screening sequence: SEQ ID NO:605 to SEQ ID NO:609.

The results shown in Figure 10A demonstrate that Distamycin A preferentially disrupts binding to the test sequences UL9 polyT, UL9 polyA and, to a lesser extent, UL9 ATAT. Figure 10A also shows the concentration dependence of the inhibitory effect of distamycin A: at 1 μ M distamycin A most of the DNA:protein complexes are intact (top band) with free probe appearing in the UL9 polyT and UL9 polyA lanes, and some free probe appearing in the UL9 ATAT lane; at 4 μ M free probe can be seen in the UL9 polyT and UL9 polyA lanes; at 16 μ M free probe can be seen in the UL9 polyT and UL9 polyA lanes; and at 40 μ M the DNA:protein in the polyT, UL9 polyA and UL9 ATAT lanes are near completely disrupted while some DNA:protein complexes in the other lanes persist. These results were consistent with the reported preference of Distamycin A for non-alternating AT-rich sequences.

Actinomycin D was tested with 8 different test sequences flanking the UL9 screening sequence: SEQ ID NO:605 to SEQ ID NO:609, and SEQ ID NO:611 to SEQ ID NO:613. The results shown in Figure 10B demonstrate that actinomycin D preferentially disrupts the binding of UL9-COOH to the oligonucleotides UL9 CCGG (SEQ ID NO:605) and UL9 GGGC (SEQ ID NO:606). These oligonucleotides contain, respectively, three or five dGdC dinucleotides in addition to the dGdC dinucleotide within the UL9 recognition sequence. This result is consistent with the results described in the literature for Actinomycin D binding to the dinucleotide sequence dGdC. Apparently the presence of a potential preferred target site within the screening sequence (*oriS*, SEQ ID NO:601), as mentioned above, does not interfere with the function of the assay.

Doxorubicin was tested with 8 different test sequences flanking the UL9 screening sequence: SEQ ID NO:605 to SEQ ID NO:609, and SEQ ID NO:611 to SEQ ID NO:613. The results shown in Figure 10C demonstrate that Doxorubicin preferentially disrupts binding to *oriEco3*, the test sequence of which differs from *oriEco2* by only one base (compare SEQ ID NO:612 and SEQ ID NO:613). Figure 10C also shows the concentration dependence of the inhibitory effect of Doxorubicin: at 15 μ M Doxorubicin, the UL9 binding to the screening sequence is strongly affected when *oriEco3* is the test sequence, and more mildly affected when polyT, UL9 GGGC, or *oriEco2* was the test sequence; and at 35 μ M Doxorubicin most DNA:protein complexes are nearly completely disrupted, with UL9 polyT and UL9ATAT showing some DNA still complexed with protein. Also, effects similar to those observed at 15 μ M were also observed using Doxorubicin at 150 nM, but at a later time point.

The feasibility studies performed with the limited set of test sequences, described above, provided evidence that the results of the assay are not inconsistent with the results reported in the literature. However, the screening of all possible 256 four base-pair sequences, using the assay of the present invention, provides a much more extensive overview of the sequence preferences of distamycin A and actinomycin D.

The actual ranking of values obtained from the assay, for any given test compound, can be variable. A number of sequences can be clustered having similar affinity: although absolute rank might not be determinable, relative ranks can be determined.

The results obtained in the feasibility studies with both distamycin A and actinomycin D were corroborated by the results obtained in the screen of all 256 sequences. In other words, the rank of the oligonucleotides remained internally consistent in the larger screen. Further, the screens of distamycin A and actinomycin D both support the general hypotheses described in the literature: that is, distamycin A has a preference for binding AT-rich sequences while actinomycin D has a preference for binding GC-rich sequences. However, both drug screens of all possible 4 bp sequences revealed additional characteristics that have not been described in the literature.

Based on the data from 4 separate experiments (Examples 10 and 11; Figures 15, 16 and 17), consensus sequences can be derived for distamycin binding. One consensus sequence (Example 11) is relatively AT-rich, although the preference in the 4th base position is distinctly G or C. The other consensus sequence (Example 11) is relatively GC-rich, with some of the sequences having a 75% GC-content. As noted above, the assay data is consistent with distamycin binding data shown in the literature.

The ability of the assay to distinguish sequence binding preference using weak DNA-binding molecules with relatively poor sequence-specificity (such as distamycin A) is a stringent test of the assay. Accordingly, the present assay seems well-suited for the identification of molecules having better sequence specificity and/or higher sequence binding affinity. Further, the results demonstrate sequence preferential binding with the known anti-cancer drug Doxorubicin. This result indicates the assay may be useful for screening mixtures for molecules displaying similar characteristics that could be subsequently tested for anti-cancer activities as well as sequence-specific binding.

Other compounds that may be suitable for testing in the present DNA:protein system or for defining alternate DNA:protein systems include the following categories of DNA-binding molecules.

A first category of DNA-binding molecules includes non-intercalating major and minor groove DNA-binding molecules. For example, two major classes of major groove binding molecules are DNA-binding proteins (or peptides) and nucleic acids (or nucleic acid analogs such as those with peptide or morpholino backbones) capable of forming triplex DNA. There are a number of non-intercalating minor groove DNA-binding molecules including, but not limited to the following: distamycin A, netropsin, mithramycin, chromomycin and oligomycin, which are used as antitumor agents and antibiotics; and synthetic antitumor agents such as berenil, phthalanilides, aromatic bisquanylhyaazones and bisquaternary ammonium heterocycles (for review, see Baguley, 1982). Non-intercalating DNA-binding molecules vary greatly in structure: for example, the netropsin-distamycin series are oligopeptides compared to the diarylamidines berenil and stilbamidine.

A second category of DNA-binding molecules includes intercalating DNA-binding molecules. Intercalating agents are an entirely different class of DNA-binding molecules that have been identified as antitumor therapeutics and include molecules such as daunomycin (Chaires, et al.) and nogalomycin (Fox, et al., 1988) (see Remers, 1984).

A third category of DNA-binding molecules includes molecules that have both groove-binding and intercalating properties. DNA-binding molecules that have both intercalating and minor groove binding properties include actinomycin D (Goodisman, et al.), echinomycin (Fox, et al. 1990), triostin A (Wang, et al.), and luzopeptin (Fox, 1988). In general, these molecules have one or two planar polycyclic moieties and one or two cyclic oligopeptides. Luzopeptins, for instance, contain two substituted quinoline chromophores linked by a cyclic decadepsiptide. They are closely related to the quinoxaline family, which includes echinomycin and triostin A, although they luzopeptins have ten amino acids in the cyclic peptide, while the quinoxaline family members have eight amino acids.

In addition to the major classes of DNA-binding molecules, there are also some small inorganic molecules, such as cobalt hexamine, which is known to induce Z-DNA formation in regions that contain repetitive GC sequences (Gessner, et al.). Another example is cisplatin, cisdiamminedichloroplatinum(II), which is a widely used anticancer therapeutic. Cisplatin forms a covalent intrastrand crosslink between the N7 atoms of adjacent guanosines (Rice, et al.).

Furthermore, there are a few molecules, such as calicheamicin, that have unusual biochemical structures that do not fall in any of the major categories. Calicheamicin is an antitumor antibiotic that cleaves DNA and is thought to recognize DNA sequences through carbohydrate moieties (Hawley, et al.). Several DNA-binding molecules, such as daunomycin, A447C, and cosmomycin B have sugar group, which may play a role in the recognition process.

Limited sequence preferences for some of the above drugs have been suggested: for example, echinomycin is thought to preferentially bind to the sequence (A/T)CGT (Fox, et al.). However, the absolute sequence preferences of the known DNA-binding drugs have never been demonstrated. Despite the large number of publications in this field, prior to the development of the assay described herein, no methods were available for determining sequence preferences among all possible binding sequences.

g. Theoretical Considerations on the Concentration of Assay Components.

There are two major components in the assay, the test oligonucleotide (i.e., the test sequence) and the DNA-binding domain of UL9, which is described below. A number of theoretical considerations have been employed in establishing the assay system. In one embodiment of the invention, the assay is used as a mass-screening assay: in this embodiment the smallest volumes and concentrations possible were desirable. Each assay typically uses about 0.1-0.5 ng DNA in a 15-20 μ l reaction volume (approximately 0.3-1.5 nM). The protein concentration is in excess and can be varied to increase or decrease the sensitivity of the assay. In the simplest scenario (stearic hindrance), where the small molecule is acting as a competitive inhibitor and the ratio of DNA:protein and DNA-binding test molecule:DNA is 1:1, the system kinetics can be described by the following equations:

$$D + P \rightleftharpoons D:P, \text{ where } k_{fp}/k_{bp} = K_{eq,p} = [D:P]/[D][P]$$

and

$$D + X \rightleftharpoons D:X, \text{ where } k_{fx}/k_{bx} = K_{eq,x} = [D:X]/[D][X]$$

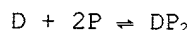
D = DNA, P = protein, X = DNA-binding molecule, k_{fp} and k_{fx} are the rates of the forward reaction for the DNA:protein interaction and DNA:drug interaction, respectively, and k_{bp} and k_{bx} are the rates of the backwards reactions for the respective interactions. Brackets, [], indicate molar concentration of the components.

In the assay, both the protein, P, and the DNA-binding molecule or drug, X, are competing for the DNA. If stearic hindrance is the mechanism of inhibition, the assumption can be made that the two molecules are competing for the same site. When the concentration of DNA equals the concentration of the DNA:drug or DNA:protein complex, the equilibrium binding constant, K_{eq} , is equal to the reciprocal of the protein concentration ($1/[P]$). When all three components are mixed together, the relationship between the drug and the protein can be described as:

$$K_{eq,p} = z(K_{eq,x})$$

where "z" defines the difference in affinity for the DNA between P and X. For example, if $z = 4$, then the affinity of the drug is 4-fold lower than the affinity of the protein for the DNA molecule. The concentration of X, therefore, must be 4-fold greater than the concentration of P, to compete equally for the DNA molecule. Thus, the equilibrium affinity constant of UL9 will define the minimum level of detection with respect to the concentration and/or affinity of the drug. Low affinity DNA-binding molecules will be detected only at high concentrations; likewise, high affinity molecules can be detected at relatively low concentrations. With certain test sequences, complete inhibition of UL9 binding at markedly lower concentrations than indicated by these analyses have been observed, probably indicating that certain sites among those chosen for feasibility studies have affinities higher than previously published. Note that relatively high concentrations of known drugs can be utilized for testing sequence specificity. In addition, the binding constant of UL9 can be readily lowered by altering the pH or salt concentration in the assay if it ever becomes desirable to screen for molecules that are found at low concentration (e.g., in a fermentation broth or extract).

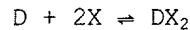
The system kinetic analysis becomes more complex if more than one protein or drug molecule is bound by each DNA molecule. As an example, if UL9 binds as a dimer,



then the affinity constant becomes dependent on the square of the protein concentration:

$$K = [DP_2]/[D][P]^2$$

The same reasoning holds true for the DNA-binding test molecule, X; if,

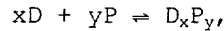


then the affinity constant becomes dependent on the square of the protein concentration:

$$K = [DX_2] / [D][X]^2$$

Similarly, if the molar ratio of DNA to DNA-binding test molecule was 1:3, the affinity constant would be dependent on the cube of the drug concentration.

Experimentally, the ratio of molar components can be determined. Given the chemical equation:



the affinity constant may be described as

$$K = [D_xP_y] / [D]^x [P]^y$$

where [] indicates concentration, D = DNA, P = protein, x = number of DNA molecules per DNA:protein complex, and y = number of protein molecules per DNA:protein complex. By determining the ratio of DNA:protein complex to free DNA, one can solve for x and y:

$$\text{if } X_{\text{total}} = X_{\text{free}} + X_{\text{bound}};$$

if a = the fraction of DNA that is free, then the fraction of DNA that is bound can be described as 1-a; and if $X_{\text{bound}}:X_{\text{free}}$ (the ratio of DNA:protein complex to free DNA) is known for more than one DNA concentration. This is because the affinity constant should not vary at different DNA concentrations. Therefore,

$$K_{D:P, [D1]} = K_{D:P, [D2]}.$$

Substituting the right side of the equation above,

$$[D1_xP_y] / [D1]^x [P]^y = [D2_xP_y] / [D2]^x [P]^y.$$

Because the concentration of components in the assay can be varied and are known, the molar ratio of the components can be determined. Therefore, $[D1_xP_y]$ and $[D2_xP_y]$ can be described as $(1-a_1)[x_1]$ and $(1-a_2)[x_2]$, respectively, and $[D1]$ and $[D2]$ can be described as $(a_1)[x_1]$ and $(a_2)[x_2]$, respectively. $[P]$ remains constant and is described as $(y)-(y/x)(1-a)(x)$, where y is the total protein concentration and $(y/x)(1-a)(x)$ is the protein complexed with DNA.

The system kinetic analyses become more complex if the inhibition is allosteric (non-competitive inhibition) rather than competition by steric hindrance. Nonetheless, the probability that the relative effect of an inhibitor on different test sequences is due to its relative and differential affinity to the different test sequences is fairly high. This is particularly true in the assays in which all sequences within an ordered set (e.g., possible sequences of a given length or all possible variations of a certain base composition and defined length) are tested. In short, if the effect of inhibition in the assay is particularly strong for a single sequence, then it is likely that the inhibitor binds that particular sequence with higher affinity than any of the other sequences. Furthermore, while it may be difficult to determine the absolute affinity of the inhibitor, the relative affinities have a high probability of being reasonably accurate. This information will be most useful in facilitating, for instance, the refinement of molecular modeling systems.

h. The Use of the Assay under Conditions of Very High Protein Concentration.

When the screening protein is added to the assay system at very high concentrations (i.e., high enough to force binding to non-specific sites -- the protein binds to non-specific sites on the oligonucleotide as well as the screening sequence). This has been demonstrated using band shift gels: when serial dilutions are made of the protein and mixed with a fixed concentration of oligonucleotide, no binding (as seen by a band shift) is observed at very low dilutions (e.g., 1:100,000), a single band shift is observed at moderate dilutions (e.g., 1:100) and a smear, migrating higher than the single band observed at moderate dilutions, is observed at high concentrations of protein (e.g., 1:10). The observation of a smear is indicative of a mixed population of complexes, all of which presumably have the screening protein binding to the screening sequence with high affinity, but in addition have a larger number of proteins bound with markedly lower affinity to other sites.

Some of the low affinity binding proteins are likely bound to the test sequence. For example, when using the UL9-based system, the low affinity binding proteins are likely UL9 or less likely glutathione-S-transferase: these are the only proteins in the assay mixture. These proteins are significantly more sensitive to interference by a molecule binding to the test sequence for two reasons. First, the interference is likely to be by direct steric hindrance and does not rely on induced conformational changes in the DNA; secondly, the protein is a low affinity binding protein because the test site is not a cognate-binding sequence. In the case of UL9, the difference in affinity between the low affinity binding and the high affinity binding appears to be at least two orders of magnitude.

The filter binding assays capture more DNA:protein complexes when more protein is bound to the DNA. The relative results are accurate, but under moderate protein concentrations, not all of the bound DNA (as demonstrated by band shift assays) will bind to the filter unless there is more than one DNA:protein complex per oligonucleotide (e.g., in the case of UL9, more than one UL9:DNA complex). This makes the assay exquisitely sensitive under conditions of high protein concentration. For instance, when actinomycin binds DNA at a test site under conditions where there is one DNA:UL9 complex per oligonucleotide, a preference for binding GC-rich oligonucleotides has been observed; under conditions of high protein concentration, where more than one DNA:UL9 complex is found per oligonucleotide, this binding preference is even more apparent. These results suggest that the effect of actinomycin D on a test site that is weakly bound by protein may be more readily detected than the effect of actinomycin D on the adjacent screening sequence. Therefore, employing high protein concentrations may increase the sensitivity of the assay.

III. Amplification-Based Selection Technique to Determine the Sequence Preferences of DNA-Binding Molecules.

A. Design of Test Oligonucleotides.

The above-described assay can be coupled to amplification methods (in one embodiment, polymerase chain reaction (Mullis, et al.; Mullis; Innis, et al.)) to achieve identification of the sequences to which binding of a test molecule is most preferred.

In this embodiment of the present invention, a double stranded test oligonucleotide is synthesized that contains the following elements:

(i) the binding site for a DNA-binding protein (for example, UL9), i.e., the screening site,

(ii) adjacent the screening site, a test site composed of more than two base pairs and preferably less than 20 base pairs (most preferably 4-12 bases), and

(iii) means to isolate selected sequences for amplification, such as a sufficient number of bases flanking the test site sequences to function as priming sites for polymerase chain reaction amplification or restriction sites useful to facilitate cloning.

Priming sites can also be used as primer binding sites for dideoxy sequencing reactions and may contain restriction endonuclease cleavage sites to facilitate cloning manipulations.

The double-stranded test oligonucleotide can be generated by second-strand synthesis using a primer complementary to the priming site at the 3' end of the top-strand of the test oligonucleotide. Alternatively, both strands can be generated by other means, such as chemical synthesis, and the double-stranded test oligonucleotides can be generated by hybridization of the strands.

An example of one such a test oligonucleotide is shown in Figure 29A (SEQ ID NO:630, SEQ ID NO:631 and SEQ ID NO:632). A specific example of a test oligonucleotide is shown in Figure 29B (SEQ ID NO:633), where X=4. All possible 256 four base pair sequences are represented at equimolar levels within the pool of oligonucleotides generated by this sequence design.

Another example of such a test oligonucleotide sequence is shown in Figure 29C (SEQ ID NO:634), for an 8 base pair test sequence. In this pool of mixed sequences, all possible 8 base pair sequences ($4^8 = 65,536$) are present in equimolar amounts.

A second set of test oligonucleotides may be constructed in which the test site is placed on the other side of the DNA-binding protein recognition site (e.g., Figure 29D, SEQ ID NO:635).

For any single-stranded test oligonucleotide pool, the single-stranded molecules are annealed to a primer and the bottom strands are enzymatically synthesized by primer extension reactions. One advantage of using the assay/amplification PCR-cycling embodiment of the present invention is that it is convenient to work with larger test sequences in this embodiment. This protocol is geared to determining the highest affinity binding sequences and is not capable of determining the rank of all test sequences nor of identifying low affinity binding sites: such ranking can be determined by screening individual sequences as described above.

B. Applying the Assay to the Mixed Pools of Test Oligonucleotides.

Using double-stranded test oligonucleotides, such as those just described, the basic assay is performed essentially as described above (Section I): typically without the use of radioactive detection systems. As previously discussed, a number of DNA:protein interactions may be used in this assay system. One example of such a system is the interaction of the DNA-binding domain of UL9 (or UL9-COOH) with its cognate recognition sequence.

In this embodiment of the present invention, UL9-COOH is added to the test oligonucleotide pool (for example, 256 four base pair sequences are represented at equimolar levels within the pool of oligonucleotides described above) in UL9 binding buffer. DNA-binding molecules are tested for the ability to differentially disrupt the binding of the UL9 DNA:protein complex by binding to the test sequence. After the addition of the test molecule or test mixture (e.g., a fermentation broth or fungal extract), the assay mixture is incubated for a desired time, then passed through a nitrocellulose filter. DNA:protein (such as DNA:UL9) complexes are captured on the filter. DNA that is not bound by protein passes through the filter (i.e., the filtrate) (step 1). The volume of the assay is adjusted to accommodate the amount required for the filtering process: that is, taking into consideration the losses incurred during the filtering process.

C. Amplification.

In one embodiment, the DNA present in the filtrate is amplified using the polymerase chain reaction (PCR) technology (Mullis; Mullis, et al.; Perkin Elmer-Cetus). An aliquot of the resulting PCR-amplified material is cycled through the DNA:protein binding assay again (step 2), then PCR-amplified again (step 3). Steps 1-3 are repeated several times using each subsequent filtrate. After each PCR amplification, part of the PCR-amplified material is retained for sequencing analysis. The result of the repeated cyclings through the assay/amplification process is that the test oligonucleotide sequences that are amplified contain test sequences that are preferred binding sites for the test molecules. Through subsequent rounds of assay/amplification, these oligonucleotides are amplified to represent a larger and larger percent of the total population of amplified DNA molecules.

In addition to PCR, the DNA present in the filtrate can be amplified by other methods as well. For example, the DNA present in the filtrate can be cloned into a selected vector (such as, phage vectors, e.g., lambda-based, or

standard cloning vectors, e.g., pBR322- or pUC-based). The cloned sequences are then transformed into an appropriate host organism in which the selected vector can replicate (for example, bacteria or yeast). The transformed host organism is cultured with concurrent amplification of the vectors containing the cloned sequences. The vectors are then isolated by standard procedures (Maniatis, et al.; Sambrook, et al.; Ausubel, et al.). Typically, the cloned sequences, originally obtained from the DNA filtrate, are obtained from the vector by restriction endonuclease digestion and size-fractionation (for example, electrophoretic separation of the digestion products followed by electroelution of the cloned sequences of interest) (Ausubel, et al.). These isolated amplified test oligonucleotide sequences can then be recycled through subsequent rounds of assay/amplification as described above.

In another embodiment, the oligonucleotide sequences present in the original DNA filtrate can be isolated, sequenced and amplified by *in vitro* synthesis of copies of the oligonucleotides.

D. Sequencing of Amplified DNA.

Samples from each cycle are sequenced using, for example, radio-labeled primers and dideoxy sequencing methodologies (Sanger) or the chemical methodologies outlined by Maxam and Gilbert. If the amplified sequences are not sufficiently resolved to obtain a unambiguous sequence information, then the DNA is further purified and sequenced. For example, the DNA is cleaved at the restriction endonuclease sites within the primer sequences and subcloned into a convenient sequencing vector, such as "BLUESCRIPT" (Stratagene, La Jolla, CA). The sequencing vectors carrying the amplified inserts are transformed into bacteria. The resulting cloned vectors are isolated and sequenced (in the case of "BLUESCRIPT," using the commercially available primers and protocols).

IV. Modifications of Test Oligonucleotides and other Useful DNA:Protein Interactions

One class of DNA:protein interactions that may be useful in the assay of the present invention is the restriction endonuclease:restriction site class of DNA:protein interactions. In the absence of divalent cations, restriction endonucleases bind DNA but have no enzymatic activity (cleavage of DNA does not take place without divalent cations). This allows the assay of the present invention to be performed using a restriction endonuclease with its cognate binding site as the screening sequence. The use of the restriction endonuclease:restriction site interaction as the basis of the present assay is described in greater detail in Section VI.B.4(c).

The test oligonucleotides of the present invention can be modified to contain two different DNA:protein screening systems, i.e., two different screening sequences with their respective cognate binding proteins. In the assay described above, the UL9 screening sequence lies on one side of and immediately adjacent to the test sequence. A second screening sequence, such as, a restriction endonuclease recognition sequence (restriction site), can be introduced immediately adjacent to the other side of the test sequence.

Several restriction enzymes may recognize the same restriction site. These enzymes are not identical, however, in that the cleavage sites may be at the 5' end, the center, or the 3' end of the recognition sequence. For this reason, a restriction site that is recognized by more than one restriction enzyme may be incorporated adjacent to the test site. This allows a single pool of test oligonucleotides to be used in assays employing three different DNA:protein interactions: the screening sequence has the same sequence but the restriction endonuclease used in the assay system is different in each case. Using this method to design test oligonucleotides, the UL9 screening sequence may be placed on one side of a test sequence and a restriction site screening sequence (having three cognate binding proteins) may be placed on the other side of the test sequence. Such a test oligonucleotide construction allows 4 different DNA:protein assay interaction systems to be employed with a single pool of test sequences.

One example of test oligonucleotides using several different DNA:protein interaction systems are shown in Figure 30. The top strands of the pool of test oligonucleotides shown in Figure 30 have 6 base pair test sequences (NNNNNN) and represent synthetic pools of all possible 4096 test sequences.

The remainder of the nucleotide sequence is fixed. The test oligonucleotides contain the UL9 recognition sequence, 5'-CGTTCGCACTT-3' (underlined) on one side of the test sequence and a restriction endonuclease binding sequence, 5'-GGTACC-3' (bold), on the other side of the test site. The restriction endonuclease recognition sequence is recognized by the three different restriction endonucleases *Asp718*, *RsaI* and *KpnI*. In Figure 30 the UL9 binding site (screening sequence) is located 3' of the test sequence: the UL9 binding site (screening sequence) can also be located 5' of the test sequence.

The shorter sequences shown above the 5' and 3' ends of the test oligonucleotides are primer sequences useful for sequencing and PCR amplification. The primer sequences contain commonly used restriction endonuclease sites for the purpose of subcloning into sequencing vectors.

Performing the assay with two or more different protein/screening sequence systems allows the confirmation of putative high affinity binding between a test compound and specific test sequences.

Alternatively, since there is no assurance that a test molecule that binds the test sequence will have the same effect on protein binding at both adjacent flanking sequences, simultaneous use of both test systems may reduce the number of false negatives detected in an assay. For example, a test molecule that does not affect the binding of protein at one screening site but may effect the binding of a different protein at the other screening site.

V. Capture/Detection Systems.

As an alternative to the above described band shift gels and filter binding assays, the measurement of inhibitors can be monitored by measuring either the level of unbound DNA in the presence of test molecules or mixtures or the level of DNA:protein complex remaining in the presence of test molecules or mixtures. Measurements may be made either at equilibrium or, in a kinetic assay, prior to the time at which equilibrium is reached. The type of measurement is likely to be dictated by practical factors, such as the length of time to equilibrium, which will be determined by both the kinetics of the DNA:protein interaction as well as the kinetics of the DNA:drug interaction. The results (i.e., the detection of DNA-binding molecules and/or the determination of their sequence preferences) should not vary with the type of measurement taken (kinetic or equilibrium).

Figure 2 illustrates an assay for detecting inhibitory molecules based on their ability to preferentially hinder the binding of a DNA-binding protein. In the presence of an inhibitory molecule (X) the equilibrium between the DNA-binding protein and its binding site (screening sequence) is disrupted. The DNA-binding protein (O) is displaced from DNA (/) in the presence of inhibitor (X), the DNA free of protein or, alternatively, the DNA:protein complexes, can then be captured and detected.

For maximum sensitivity, unbound DNA and DNA:protein complexes should be sequestered from each other in an efficient and rapid manner. The method of DNA capture should allow for the rapid removal of the unbound DNA from the protein-rich mixture containing the DNA:protein complexes.

Even if the test molecules are specific in their interaction with DNA they may have relatively low affinity and they may also be weak binders of non-specific DNA or have non-specific interactions with DNA at low concentrations.

In either case, their binding to DNA may only be transient, much like the transient binding of the protein in solution. Accordingly, one feature of the assay is to take a molecular snapshot of the equilibrium state of a solution comprised of the test oligonucleotide DNA, the protein, and the inhibitory test molecule. In the presence of an inhibitor, the amount of DNA that is not bound to protein will be greater than in the absence of an inhibitor. Likewise, in the presence of an inhibitor, the amount of DNA that is bound to protein will be lesser than in the absence of an inhibitor.

Any method used to separate the DNA:protein complexes from unbound DNA, should be rapid, because when the capture system is applied to the solution (if the capture system is irreversible), the ratio of unbound DNA to DNA:protein complex will change at a predetermined rate, based purely on the off-rate of the DNA:protein complex. This step, therefore, determines the limits of background. Unlike the protein and inhibitor, the capture system should bind rapidly and tightly to the DNA or DNA:protein complex. The longer the capture system is left in contact with the entire mixture of unbound DNA and

DNA:protein complexes in solution, the higher the background, regardless of the presence or absence of inhibitor.

Two exemplary capture systems are described below for use in the assay of the present invention. One capture system has been devised to capture unbound DNA (Section V.A). The other has been devised to capture DNA:protein complexes (Section V.B). Both systems are amenable to high throughput screening assays.

The same detection methods (Section V.C) can be applied to molecules captured using either capture system.

10 A. Capture of Unbound DNA.

One capture system that has been developed in the course of experiments performed in support of the present invention utilizes a streptavidin/biotin interaction for the rapid capture of unbound DNA from the protein-rich mixture, which includes unbound DNA, DNA:protein complexes, excess protein and the test molecules or test mixtures. Streptavidin binds with extremely high affinity to biotin ($K_d = 10^{-15}M$) (Chalet, *et al.*; Green). Accordingly, two advantages of the streptavidin/biotin system are that binding between the two molecules can be rapid and the interaction is the strongest known non-covalent interaction.

In this detection system a biotin molecule is covalently attached in the oligonucleotide screening sequence (*i.e.*, the DNA-binding protein's binding site). This attachment is accomplished in such a manner that the binding of the DNA-binding protein to the DNA is not destroyed. Further, when the protein is bound to the biotinylated sequence, the protein prevents the binding of streptavidin to the biotin. In other words, the DNA-binding protein is able to protect the biotin from being recognized by the streptavidin. This DNA:protein interaction is illustrated in Figure 3.

The capture system is described herein for use with the UL9/*oriS* system described above. The following general testing principles can, however, be applied to analysis of other DNA:protein interactions. The usefulness of this system depends on the biophysical characteristics of the particular DNA:protein interaction.

1. Modification of the Protein Recognition Sequence with Biotin.

The recognition sequence for the binding of the UL9 (Koff, *et al.*) protein is underlined in Figure 4. Oligonucleotides were synthesized that contain the UL9 binding site and site-specifically biotinylated a number of locations throughout the binding sequence (SEQ ID NO:614; Example 1, Figure 4).

These biotinylated oligonucleotides were then used in band shift assays to determine the ability of the UL9 protein to bind to the oligonucleotide. These experiments using the biotinylated probe and a non-biotinylated probe as a control demonstrate that the presence of a biotin at the #8-T (biotinylated deoxyuridine) position of the bottom strand meets the requirements listed above: the presence of a biotin moiety at the #8 position of the bottom strand does not markedly affect the specificity of UL9 for the recognition site.

Further, in the presence of bound UL9, streptavidin does not recognize the presence of the biotin moiety in the oligonucleotide. Biotinylation at other A or T positions did not have the two necessary characteristics (*i.e.*, UL9 binding and protection from streptavidin): biotinylation at the adenosine in position #8, of the top strand, prevented the binding of UL9; biotinylation of either adenosines or thymidines (top or bottom strand) at positions #3, #4, #10, or #11 all allowed binding of UL9, but in each case, streptavidin also was able to recognize the presence of the biotin moiety and thereby bind the oligonucleotide in the presence of UL9.

The above result (the ability of UL9 to bind to an oligonucleotide containing a biotin within the recognition sequence and to protect the biotin from streptavidin) was unexpected in that methylation interference data (Koff, *et al.*) suggest that methylation of the deoxyguanosine residues at positions #7 and #9 of the recognition sequence (on either side of the biotinylated deoxyuridine) blocks UL9 binding. In these methylation interference experiments, guanosines are methylated by dimethyl sulfate at the N⁷ position, which corresponds structurally to the 5-position of the pyrimidine ring at which the deoxyuridine is biotinylated. These moieties all protrude into the major groove of the DNA. The methylation interference data suggest that the #7 and #9 position deoxyguanosines are contact points for UL9, it was therefore

unexpected that the presence of a biotin moiety between them would not interfere with binding.

The binding of the full length protein was relatively unaffected by the presence of a biotin at position #8 within the UL9 binding site. The rate of dissociation was similar for full length UL9 with both biotinylated and unbiotinylated oligonucleotides. However, the rate of dissociation of the truncated UL9-COOH polypeptide was faster with the biotinylated oligonucleotides than with non-biotinylated oligonucleotides (for non-biotinylated oligonucleotides the rate comparable to that of the full length protein with either DNA).

The binding conditions were optimized for UL9-COOH so that the half-life of the truncated UL9 from the biotinylated oligonucleotide was 5-10 minutes (optimized conditions are given in Example 4), a rate compatible with a mass screening assay. The use of multi-well plates to conduct the DNA:protein assay of the present invention is one approach to mass screening.

2. Capture of Site-Specific Biotinylated Oligonucleotides.

The streptavidin:biotin interaction can be employed in several different ways to remove unbound DNA from the solution containing the DNA, protein, and test molecule or mixture. Magnetic polystyrene or agarose beads, to which streptavidin is covalently attached or attached through a covalently attached biotin, can be exposed to the solution for a brief period, then removed by use, respectively, of magnets or a filter mesh. Magnetic streptavidinated beads are currently the method of choice. Streptavidin has been used in many of these experiments, but avidin is equally useful.

An example of a second method for the removal of unbound DNA is to attach streptavidin to a filter by first linking biotin to the filter, binding streptavidin, then blocking nonspecific protein binding sites on the filter with a nonspecific protein such as albumin. The mixture is then passed through the filter, unbound DNA is captured and the bound DNA passes through the filter. This method can give high background due to partial retention of the DNA:protein complex on the filter.

One convenient method to sequester captured DNA is the use of streptavidin-conjugated superparamagnetic polystyrene beads as described in Example 7. These beads are added to the assay mixture to capture the unbound DNA. After capture of DNA, the beads can be retrieved by placing the reaction tubes in a magnetic rack, which sequesters the beads on the reaction chamber wall while the assay mixture is removed and the beads are washed. The captured DNA is then detected using one of several DNA detection systems, as described below.

Alternatively, avidin-coated agarose beads can be used. Biotinylated agarose beads (immobilized D-biotin, Pierce) are bound to avidin. Avidin, like streptavidin, has four binding sites for biotin. One of these binding sites is used to bind the avidin to the biotin that is coupled to the agarose beads via a 16 atom spacer arm: the other biotin binding sites remain available. The beads are mixed with binding mixtures to capture biotinylated DNA (Example 7).

Alternative methods (Harlow, et al.) to the bead capture methods just described include the following streptavidinated or avidinated supports: low-protein binding filters, or 96-well plates.

B. Capture of DNA:Protein Complexes.

The amount of DNA:protein complex remaining in the assay mixture in the presence of an inhibitory molecule can also be determined as a measure of the relative effect of the inhibitory molecule. A net decrease in the amount of DNA:protein complex in response to a test molecule is an indication of the presence of an inhibitor. DNA molecules that are bound to protein can be captured on nitrocellulose filters. Under low salt conditions, DNA that is not bound to protein freely passes through the filter. Thus, by passing the assay mixture rapidly through a nitrocellulose filter, the DNA:protein complexes and unbound DNA molecules can be rapidly separated. This has been accomplished on nitrocellulose discs using a vacuum filter apparatus or on slot blot or dot blot apparatuses (all of which are available from Schleicher and Schuell, Keene, NH). The assay mixture is applied to and rapidly passes through the wetted nitrocellulose under vacuum conditions. Any apparatus employing nitrocellulose filters or other filters capable of retaining protein while

allowing free DNA to pass through the filter would be suitable for this system.

C. Detection Systems.

For either of the above capture methods, the amount of DNA that has been captured is quantitated. The method of quantitation depends on how the DNA has been prepared. If the DNA is radioactively labelled, beads can be counted in a scintillation counter, or autoradiographs can be taken of dried gels or nitrocellulose filters. The amount of DNA has been quantitated in the latter case by a densitometer (Molecular Dynamics, Sunnyvale, CA); alternatively, filters or gels containing radiolabeled samples can be quantitated using a phosphorimager (Molecular Dynamics). Further, the captured DNA may be detected using a chemiluminescent or colorimetric detection system.

Radiolabelling and chemiluminescence (i) are very sensitive, allowing the detection of sub-femtomole quantities of oligonucleotide, and (ii) use well-established techniques. In the case of chemiluminescent detection, protocols have been devised to accommodate the requirements of a mass-screening assay. Non-isotopic DNA detection techniques have principally incorporated alkaline phosphatase as the detectable label given the ability of the enzyme to give a high turnover of substrate to product and the availability of substrates that yield chemiluminescent or colored products.

1. Radioactive Labeling.

Many of the experiments described above for UL9 DNA:protein-binding studies have made use of radio-labelled oligonucleotides. The techniques involved in radiolabelling of oligonucleotides have been discussed above. A specific activity of 10^8 - 10^9 dpm per μ g DNA is routinely achieved using standard methods (e.g., end-labeling the oligonucleotide with adenosine γ -[32 P]-5' triphosphate and T4 polynucleotide kinase). This level of specific activity allows small amounts of DNA to be measured either by autoradiography of gels or filters exposed to film or by direct counting of samples in scintillation fluid.

2. Chemiluminescent Detection.

For chemiluminescent detection, digoxigenin-labelled oligonucleotides (Example 1) can be detected using the chemiluminescent detection system "SOUTHERN LIGHTS," developed by Tropix, Inc. (Bedford, MA). The detection system is diagrammed in Figures 11A and 11B. The technique can be applied to detect DNA that has been captured on either beads, filters, or in solution.

Alkaline phosphatase is coupled to the captured DNA without interfering with the capture system. To do this several methods, derived from commonly used ELISA (Harlow, et al.; Pierce, Rockford IL) techniques, can be employed. For example, an antigenic moiety is incorporated into the DNA at sites that will not interfere with (i) the DNA:protein interaction, (ii) the DNA:drug interaction, or (iii) the capture system. In the UL9 DNA:protein/biotin system the DNA has been end-labelled with digoxigenin-11-dUTP (dig-dUTP) and terminal transferase (Example 1, Figure 4). After the DNA was captured and removed from the DNA:protein mixture, an anti-digoxigenin-alkaline phosphatase conjugated antibody was then reacted (Boehringer Mannheim, Indianapolis IN) with the digoxigenin-containing oligonucleotide. The antigenic digoxigenin moiety was recognized by the antibody-enzyme conjugate. The presence of dig-dUTP altered neither the ability of UL9-COOH protein to bind the *oriS* (SEQ ID NO:601)-containing DNA nor the ability of streptavidin to bind the incorporated biotin.

Captured DNA was detected using the alkaline phosphatase-conjugated antibodies to digoxigenin as follows. One chemiluminescent substrate for alkaline phosphatase is 3-(2'-spiroadamantane)-4-methoxy-4-(3"-phosphoryloxy) phenyl-1,2-dioxetane disodium salt (AMPPD) (Example 7). Dephosphorylation of AMPPD results in an unstable compound, which decomposes, releasing a prolonged, steady emission of light at 477 nm. Light measurement is very sensitive and can detect minute quantities of DNA (e.g., 10^2 - 10^3 attomoles) (Example 7).

Colorimetric substrates for the alkaline phosphatase system have also been tested. While the colorimetric substrates are useable in the present assay system, use of the light emission system is more sensitive.

An alternative to the above biotin capture system is to use digoxigenin

in place of biotin to modify the oligonucleotide at a site protected by the DNA-binding protein at the assay site: biotin is then used to replace the digoxigenin moieties in the above described detection system. In this arrangement the anti-digoxigenin antibody is used to capture the oligonucleotide probe when it is free of bound protein. Streptavidin conjugated to alkaline phosphatase is then used to detect the presence of captured oligonucleotides.

D. Alternative Methods for Detecting Molecules that Increase the Affinity of the DNA-Binding Protein for its Cognate Site.

In addition to identifying molecules or compounds that cause a decreased affinity of the DNA-binding protein for the screening sequence, molecules may be identified that increase the affinity of the protein for its cognate binding site. In this case, leaving the capture system for unbound DNA in contact with the assay for increasing amounts of time allows the establishment of a fixed half-life for the DNA:protein complex (for example, using SEQ ID NO:601/UL9). In the presence of a stabilizing molecule, the half-life, as detected by the capture system time points, will be shortened.

Using the capture system for DNA:protein complexes to detect molecules that increase the affinity of the DNA-binding protein for the screening sequence requires that an excess of unlabeled oligonucleotide containing the UL9 binding site (but not the test sequences) is added to the assay mixture. This is, in effect, an off-rate experiment. In this case, the control sample (no test molecules or mixtures added) will show a fixed off-rate. For example, samples would be taken at fixed intervals after the addition of the unlabeled competition DNA molecule, applied to nitrocellulose, and a decreasing amount of radiolabeled DNA:protein complex would be observed). In the presence of a DNA-binding test molecule that enhanced the binding of UL9, the off-rate would be decreased (i.e., the amount of radiolabeled DNA:protein complexes observed would not decrease as rapidly at the fixed time points as in the control sample).

VI. Utility.

A. The Usefulness of Sequence-Specific DNA-Binding Molecules.

The present invention defines a high through-put *in vitro* screening assay to test large libraries of biological or chemical mixtures for the presence of DNA-binding molecules having sequence binding preference. The assay is also capable of determining the sequence-specificity and relative affinity of known DNA-binding molecules or purified unknown DNA-binding molecules. Sequence-specific DNA-binding molecules are of particular interest for several reasons, which are listed here. These reasons, in part, outline the rationale for determining the usefulness of DNA-binding molecules as therapeutic agents:

First, for a given DNA:protein interaction, there are generally several thousands fewer target DNA-binding sequences per cell than protein molecules that bind to the DNA. Accordingly, even fairly toxic molecules might be delivered in sufficiently low concentration to exert a biological effect by binding to the target DNA sequences.

Second, DNA has a relatively more well-defined structure compared to RNA or protein. Since the general structure of DNA has less tertiary structural variation, identifying or designing specific binding molecules should be easier for DNA than for either RNA or protein. Double-stranded DNA is a repeating structure of deoxyribonucleotides that stack atop one another to form a linear helical structure. In this manner, DNA has a regularly repeating "lattice" structure that makes it particularly amenable to molecular modeling refinements and hence, drug design and development.

Third, since many single genes (i.e., genes which have only 1 or 2 copies in the cell) are transcribed into more than one, potentially as many as thousands of RNA molecules, each of which may be translated into many proteins, targeting any DNA site, whether it is a regulatory sequence, non-coding sequence or a coding sequence, may require a much lower drug dose than targeting RNAs or proteins. Proteins (e.g., enzymes, receptors, or structural proteins) are currently the targets of most therapeutic agents. More recently, RNA molecules have become the targets for antisense or ribozyme therapeutic molecules.

Fourth, blocking the function of a RNA that encodes a protein or of the protein itself when that protein regulates several cellular genes may have

detrimental effects: particularly if some of the regulated genes are important for the survival of the cell. However, blocking a DNA-binding site that is specific to a single gene regulated by such a protein results in reduced toxicity.

5 An example situation is HNF-1 binding to Hepatitis B virus (HBV): HNF-1 binds an HBV enhancer sequence and stimulates transcription of HBV genes (Chang, et al.). In a normal cell HNF-1 is a nuclear protein that appears to be important for the regulation of many genes, particularly liver-specific genes (Courtois, et al.). If molecules were isolated that specifically bound
10 to the DNA-binding domain of HNF-1, all of the genes regulated by HNF-1 would be down-regulated, including both viral and cellular genes. Such a drug could be lethal since many of the genes regulated by HNF-1 may be necessary for liver function. However, the assay of the present invention presents the ability to screen for a molecule that could distinguish the HNF-1 binding region of the
15 Hepatitis B virus DNA from cellular HNF-1 sites by, for example, including divergent flanking sequences when screening for the molecule. Such a molecule would specifically block HBV expression without effecting cellular gene expression.

20 B. General Applications of the Assay.

General applications of the assay include but are not limited to: screening libraries of unknown chemicals, either biological or synthetic compounds, for sequence-specific DNA-binding molecules, determining the sequence-specificity or preference and/or relative affinities of DNA-binding
25 molecules, testing of modified derivatives of DNA-binding molecules for altered specificity or affinity, using the assay in secondary confirmatory or mechanistic experiments, using the data generated from the above applications to refine the predictive capabilities of molecular modeling systems, and using the refined molecular modeling systems to generate a new "alphabet" of DNA-binding subunits that can be polymerized to make novel heteropolymers designed
30 de novo to bind specific DNA target sites.

35 1. Mass-Screening of Libraries for the Presence of Sequence-Specific DNA-Binding Molecules.

Many organizations (e.g., the National Institutes of Health, pharmaceutical and chemical corporations) have large libraries of chemical or biological compounds from synthetic processes or fermentation broths or extracts that may contain as yet unidentified DNA-binding molecules. One
40 utility of the assay is to apply the assay system to the mass-screening of these libraries of different broths, extracts, or mixtures to detect the specific samples that contain the DNA-binding molecules. Once the specific mixtures that contain the DNA-binding molecules have been identified, the assay has a further usefulness in aiding in the purification of the DNA-binding molecule from the crude mixture. As purification schemes are applied to the
45 mixture, the assay can be used to test the fractions for DNA-binding activity.

The assay is amenable to high throughput (e.g., a 96-well plate format automated on robotics equipment such as a Beckman Biomek workstation [Beckman, Palo Alto, CA] with detection using semi-automated plate-reading densitometers, luminometers, or phosphorimagers).

50 The concentration of protein used in mass-screening is determined by the sensitivity desired. The screening of known compounds, as described in Section VI.B.2, is typically performed in protein excess at a protein concentration high enough to produce 90-95% of the DNA bound in DNA:protein complex. The assay is very sensitive to discriminatory inhibition at this protein
55 concentration. For some mass-screening, it may be desirable to operate the assay under higher protein concentration, thus decreasing the sensitivity of the assay so that only fairly high affinity molecules will be detected: for example, when screening fermentation broths with the intent of identifying high affinity binding molecules. The range of sensitivities in the assay will be
60 determined by the absolute concentration of protein used.

One utility of the method of the present invention, under conditions using a relatively insensitive system (high [P]:[D] ratio), is as a screening system for novel restriction enzymes. In this case, an ability to discriminate
65 between slight differences in affinity to different sequences may not be necessary or desirable. Restriction enzymes have highly discriminatory

recognition properties -- the affinity constant of a restriction endonuclease for its specific recognition sequence versus non-specific sequences are orders of magnitude different from one another. The assay may be used to screen bacterial extracts for the presence of novel restriction endonucleases. The 256 test oligonucleotides described in Example 10, for example, may be used to screen for novel restriction endonucleases with 4 bp recognition sequences. The advantages of the system are that all possible 4 bp sequences are screened simultaneously, that is, it is not limited to self-complementary sequences. Further, any lack of specificity (such as, more than one binding site) is uncovered during the primary screening assay.

2. Directed Screening.

The assay of the present invention is also useful for screening molecules that are currently described in the literature as DNA-binding molecules but with uncertain DNA-binding sequence specificity (i.e., having either no well-defined preference for binding to specific DNA sequences or having certain higher affinity binding sites but without defining the relative preference for all possible DNA binding sequences). The assay can be used to determine the specific binding sites for DNA-binding molecules, among all possible choices of sequence that bind with high, low, or moderate affinity to the DNA-binding molecule. Actinomycin D, Distamycin A, and Doxorubicin (Example 6) all provide examples of molecules with these modes of binding. Many anti-cancer drugs, such as Doxorubicin (see Example 6), show binding preference for certain identified DNA sequences, although the absolute highest and lowest specificity sequences have yet to be determined, because, until the invention described herein, methods (Salas and Portugal; Cullinane and Phillips; Phillips; and Phillips, et al.) for detecting differential affinity DNA-binding sites for any drug were limited. Doxorubicin is one of the most widely used anti-cancer drugs currently available. As shown in Example 6, Doxorubicin is known to bind some sequences preferentially. Another example of such sequence binding preference is Daunorubicin (Chen, et al.) which differs slightly in structure from Doxorubicin (Goodman, et al.). Both Daunorubicin and Doxorubicin are members of the anthracycline antibiotic family: antibiotics in this family, and their derivatives, are among the most important newer antitumor agents (Goodman, et al.).

The assay of the present invention allows the sequence preferences or specificities of DNA-binding molecules to be determined. The DNA-binding molecules for which sequence preference or specificity can be determined may include small molecules such as aminoacridines and polycyclic hydrocarbons, planar dyes, various DNA-binding antibiotics and anticancer drugs, as well as DNA-binding macromolecules, such as, peptides and polymers that bind to nucleic acids (e.g., DNA and the derivatized homologs of DNA that bind to the DNA helix).

The molecules that can be tested in the assay for sequence preference/specificity and relative affinity to different DNA sites include both major and minor groove binding molecules as well as intercalating and non-intercalating DNA binding molecules.

3. Molecules Derived from Known DNA-binding Molecules.

The assay of the present invention facilitates the identification of different binding activities by molecules derived from known DNA-binding molecules. An example of this would be to identify and test derivatives of anti-cancer drugs that have DNA-binding activity and then test for anti-cancer activity through, for example, a battery of assays performed by the National Cancer Institute (Bethesda MD). Further, the assay of the present invention can be used to test derivatives of known anti-cancer agents to examine the effect of the modifications on DNA-binding activity and specificity. In this manner, the assay may reveal activities of anti-cancer agents, and derivatives of these agents, that facilitate the design of DNA-binding molecules with therapeutic or diagnostic applications in different fields, such as antiviral or antimicrobial therapeutics. The binding-activity information for any DNA-binding molecule, obtained by application of the present assay, can lead to a better understanding of the mode of action of more effective therapeutics.

4. Secondary Assays.

As described above, the assay of the present invention is used (i) as a screening assay to detect novel DNA-binding molecules, or (ii) to determine the relative specificity and affinity of known molecules (or their derivatives). The assay may also be used in confirmatory studies or studies to elucidate the binding characteristics of DNA-binding molecules. Using the assay as a tool for secondary studies can be of significant importance to the design of novel DNA-binding molecules with altered or enhanced binding specificities and affinities.

a.) Confirmatory Studies.

The assay of the present invention can be used in competition studies to confirm and refine the original direct binding data obtained from the assay.

The primary screening assay does not provide for the direct determination of relative absolute affinities of test molecules for different test sequences.

A competition method has been developed that aids in the interpretation and confirmation of the primary screening assay. The competition method also provides a means for determining the minimum difference in absolute affinities of any test sequences for a given test molecule.

Sequences of interest are tested for their ability to compete with the test oligonucleotide for binding a test molecule of interest. In this method, DNA molecules that contain sequences that are high affinity binding sites for the DNA-binding test molecule compete effectively with the test oligonucleotide for the binding of the test molecule. DNA molecules that contain sequences that are low affinity binding sites for the test molecules are ineffective competitors. In effect, the fold-difference in concentration required between a high affinity competitor DNA and a low affinity competitor DNA, where the competitor is required to compete with the test oligonucleotide for the binding of the DNA-binding test molecule, should be proportional to the difference in affinity between the two competitor DNA molecules.

Any test oligonucleotide may be used in the competition study. However, in practice, since most secondary screening will be used to examine the putative high affinity binding test sequences, the secondary competition assay is typically used to test a competitor oligonucleotide which is a putative high affinity test sequence.

In the competition assay, the assay conditions are essentially the same as the conditions used in the primary screening assay. The assay components are mixed, with the exception of the DNA. The mixture includes protein, buffer and the DNA-binding test molecule (control samples lack the test molecule). A test oligonucleotide is labeled (for example, using a radioisotope, although any of the described capture/detection systems should be effective in the competition study). The DNA sample, including the radiolabeled test oligonucleotide and unlabelled competitor DNA is added to the assay mixture. Typically, the competitor DNA of interest is added to different reactions over a range of competitor concentrations. Two controls are commonly run: (i) no DNA binding test molecule added; and (ii) test DNA but no competitor DNA added.

The reactions are incubated for the desired time and the DNA:protein complexes separated from free DNA (i.e., DNA not associated with protein) by passing the mixture through nitrocellulose. Other capture systems, such as the biotin/streptavidin system discussed in Section V, are also effective. The amount of radiolabeled test oligonucleotide bound by protein (i.e., bound to the filter) is indicative of the effect of the competitor.

One example of a competition assay is as follows. A test oligonucleotide containing the test sequence TTAC ranks as a high affinity binding site for a test molecule. The TTAC test oligonucleotide is radiolabeled and mixed with non-radiolabeled competitor DNAs that contain, for example, a putative high affinity binding site (the same site, TTAC, is one example) or a putative low affinity binding site (e.g., CCCC). In the absence of any competing nonlabeled DNA or DNA-binding test molecule, the amount of radiolabeled DNA:protein complex observed (called $r\%$) is arbitrarily established as 100%. The concentration of the protein used in this experiment is high enough to bind most of the radiolabelled test oligonucleotide in the absence of test molecules or competing DNA molecules (this is essentially the same concentration as used in the primary screening assay).

The test molecule is added to the reaction at a concentration sufficient

to markedly reduce $r\%$, the amount of observed DNA:protein complex. The greater the reduction in signal, the more easily competition is observed. The amount of competitor DNA needed to observe competition is proportional to the amount of DNA-binding test molecule used; therefore, the amount of test molecule used should be sufficient to reduce $r\%$ to between approximately 10% to 70%. The effect of an effective competitor, such as TTAC, is to cause $r\%$ to rise towards 100%.

The competition for test molecule binding is between the non-labeled competitor DNA and the radiolabeled test oligonucleotide. As the competitor DNA concentration increases, the test molecule binds to the competitor DNA and is effectively removed from solution. Accordingly, the test molecule is no longer able to block the binding of the protein to the radiolabeled oligonucleotide. A less effective competitor, typically a competitor DNA with low affinity for the test molecule, will compete less effectively for the DNA-binding test molecule, even at substantially higher concentrations than the high affinity competitor. A completely ineffective competitor, i.e., one that did not bind the test molecule, would not cause the $r\%$ value to change, even at high concentrations of the competitor DNA.

When a competitor DNA has some affinity for the test molecule, competition ($r\%$ rising towards 100%) would be observed at some competitor DNA concentration. The difference in concentration between two competing DNA sequences to achieve an equivalent $r\%$ (e.g., 90%) should reflect the relative difference in absolute affinity between the two competitor DNA molecules. For example, if 5 μM TTAC is required to achieve a change in $r\%$ from 50% to 90% in the presence of a test molecule and 200 μM CCCC is required to achieve the same change in $r\%$, then the fold difference in affinity between TTAC and CCCC for the test molecule is $200/5 = 40$ -fold.

In the context of screening distamycin with all possible 256 bp test sequences (Example 10), the confirmatory assay can be used (i) to confirm the rankings observed in the assay, (ii) to refine the rankings among the 5-10 highest ranked binders (which show no statistical difference in rank with data from 4 experiments), and (iii) to resolve perceived discrepancies in the assay data. All of these goals may be accomplished using a competition experiment which determines the relative ability of test sequences to compete for the binding of distamycin.

The perceived discrepancy in the distamycin experiment is as follows: test oligonucleotides scored poorly in the assay which were complementary to most of the top-ranking test sequence oligonucleotides (Examples 10 and 11). This result was unexpected since it is unlikely that the affinity of distamycin for binding a test site depends on the orientation of the screening site to the test site. More likely, the assay detects the binding of distamycin when the molecule is bound to the test oligonucleotide in one orientation, but fails to detect the binding of distamycin when the test sequence is in the other orientation. A competition study will resolve this question, since the binding of distamycin to a competitor sequence will be orientation-independent; the competition does not depend on the mechanism of the assay.

For the competition experiment, the assay may be performed under any conditions suitable for the detection of drug binding. When these conditions are established, different competitor DNAs are added to the assay system to determine their relative ability to compete for drug binding with the radiolabeled test oligonucleotide in the assay system.

The competitor DNAs may be any sequence of interest. Several classes of DNA may be tested as competitor molecules including, but not limited to, the following: genomic DNAs, synthetic DNAs (e.g., poly(dA), poly(dI-dC), and other DNA polymers), test oligonucleotides of varying sequences, or any molecule of interest that is thought to compete for distamycin binding.

When using the competition assay to verify the results of a 256 oligonucleotide panel screen (like Example 10), the following criteria are useful for selecting the competitor test oligonucleotides:

(i) sequences that rank high in the assay but which do not have relative binding affinities with differences that are statistically significant from each other, in order to determine their relative affinity with greater precision;

(ii) sequences that are purported by other techniques (e.g., footprinting

or transcriptional block analysis) to be high affinity binding sites, in order to compare the results of those techniques with the screening assay results;

(iii) sequences that are complementary to test sequences that rank high in the assay, in order to determine whether these test sequences are false negatives; and

(iv) sequences of any rank in the assay, in order to confirm the assay results.

Several methods may be used to perform the competition study as long as the relative affinities of the competing DNA molecules are detectable. One such method is described in Example 14. In this example, the concentration of the assay components (drug, protein, and DNA) is held constant relative to those used in the original screening assay, but the molar ratio of the test oligonucleotide to the competitor oligonucleotides is varied.

Another method for performing a competition assay is to hold the concentrations of protein, drug and initial amount of test oligonucleotide constant, then add a variable concentration of competitor DNA. In this design, the protein and drug concentration must be sufficiently high to allow the addition of further competitor DNA without i) decreasing the amount of DNA:protein complex in the absence of drug to a level that is unsuitable for detection of DNA:protein complex, and ii) increasing the amount of DNA:protein complex in the presence of drug to a level that is unsuitable for the detection of drug binding. The window between detectable DNA:protein complex and detectable effect of the drug must be wide enough to determine differences among competitor DNAs.

In any competition method, it is important that the relative concentrations of the competing DNA molecules are accurately determined. One method for accomplishing accurate determination of the relative concentrations of the DNA molecules is to tracer-label competitor molecules to a low specific activity with a common radiolabeled primer (Example 14). In this manner, the competitor molecules have the same specific activity, but are not sufficiently radioactive (200-fold less than the test oligonucleotide) to contribute to the overall radioactivity in the assay.

b.) Secondary Studies to Elucidate Binding Characteristics.

The studies outlined in Section VI.B.4.a describe methods of determining some of the binding processes of distamycin A. The assay of the present invention may also be used to explore mechanistic questions about distamycin binding.

For example, several of the complements of the putative high affinity binding sites for distamycin have low scores in the assay. As described above, this may imply directionality in binding. The results may also imply that the test sites are not equal with respect to the effect exerted on UL9-COOH binding. Oligonucleotides can be designed to test the hypothesis of directionality.

The basic test oligonucleotide has the structure presented in Figure 27A (SEQ ID NO:621). In one scenario, the score in the binding assay is high, i.e., the greatest effect of distamycin, when the test sequences is XYZZ (Figure 27A, with the base X complementary to the base Y and the base Q complementary to the base Z), and the complement (Figure 27B; SEQ ID NO:622) scores low. These results imply that the test sites are not equivalent with respect to their effect on UL9, otherwise the right side would have the effect in one oligonucleotide and the left site would have the effect in the other. These results further suggest that the effect of distamycin is directional. The only assumption is that distamycin should bind with the same affinity to the XYZZ/QQXY sequence (Figures 27A and 27B) regardless of its position or orientation in the oligonucleotide. Since the scores are derived at equilibrium, this is likely to be the case.

To test the hypothesis that one site is effective in the assay, oligonucleotides may be designed that have the UL9 site inverted with respect to the test sites (Figures 27C and 27D; SEQ ID NO:623 and SEQ ID NO:624, respectively). If only one site is active with respect to UL9 and if the Figure 27A oligo was most effective in binding distamycin, then the oligo C should be less active in the assay than oligo D; in other words, flipping the UL9 site will result in QQXY ranking high, XYZZ ranking low.

Finally, to determine the "direction" of distamycin binding, mix test se-

quences and invert the binding site as shown in the four oligonucleotides presented in Figures 27E, 27F, 27G and 27H. Alternatively, one test site or the other could be deleted from the test oligonucleotide.

5 This type of analysis provides an example of the usefulness in the assay in determining binding properties of DNA-binding drugs.

c.) Restriction Endonucleases as Indicator Proteins in the Assay. Other DNA:protein interactions that are useful as screening sequences and their cognate binding proteins (indicator proteins) are restriction enzymes. Such secondary screening assays are performed using the same criteria to establish conditions for the primary screening assay (described in Example 4). The assay conditions can be varied to accommodate different DNA:protein interactions, as long as the assay system follows the functional criteria discussed above (Section I).

15 One limitation of using restriction endonucleases in the method of the present invention is that the assay buffer should not contain divalent cations.

In the absence of divalent cations, the enzymes will bind the appropriate recognition sequence, but not cleave the DNA. In the presence of divalent cations, the test oligonucleotide can be cleaved at or near the protein binding site.

20 By using different indicator proteins, a different recognition sequence can be used to flank the test site. This variation allows the resolution of questions regarding the potential binding of a test molecule to a site internal to any single screening sequence. For example, the assay system is used where the UL9 protein and its recognition sequence are used as the indicator protein:screening sequence interaction. In this system, if the highest affinity binding site for a test molecule is TTAC, then several test sequences may be predicted to rank high in the assay system: several of these test sequences are presented in Figure 31. In Figure 31, the test site is shown in bold, the potential binding site for the test molecule is shown underlined.

25 One test oligonucleotide on which the DNA-binding test molecule would be predicted to have a high level of effect is the oligonucleotide containing the test site, TTAC (Figure 31). However, since the UL9 recognition sequence contains the sequence TT, flanking the test site, several other test oligonucleotides might also be expected to have high activity in the assay (see Figure 31).

30 By using a different DNA:protein interaction as the indicator system in a secondary screening assay, the "false positives" shown for TACN and ACNN (shown in Figure 31) can be identified. The recognition sequence for the protein in a secondary screening assay simply needs to have a different screening sequence in the region flanking the test site than the UL9 screening sequence.

35 Restriction endonucleases provide an entire class of different DNA:protein interactions with a wide array of available sequences that can be used in this manner. For example, *SmaI* recognizes the sequence 5'-CCCGGG-3'. Using the *SmaI*:DNA interaction and the same test sequences presented in Figure 31, the resulting test oligonucleotides would have the test sequences presented in Figure 32. As can be seen from a comparison of Figures 31 and 32, changing the screening sequence from the UL9-binding sequence to the *SmaI*-binding sequence eliminates the potential test molecule binding sites internal to the screening sequence (e.g., compare TACN and ACNN in the figures).

50 The use of different DNA-binding proteins as indicator proteins in the assay is also applicable to the PCR-based test oligonucleotide selection technology (Section III).

55 5. Generation of Binding Data and Refinement of Molecular Modeling Systems.

The assay of the present invention generates data which can be applied to the refinement of molecular modeling systems that address DNA structural analysis: the data is also useful in the design and/or refinement of DNA-binding drugs. Traditionally, mass screening has been the only reasonable method for discovering new drugs. Modern rational drug design seeks to minimize laboratory screening. However, *ab initio* rational drug design is difficult at this time given (i) insufficiencies in the underlying theories used for *de novo* design, and (ii) the computational intensity which accompanies

such design approaches.

The *ab initio* approach requires calculations from first principles by quantum mechanics: such an approach is expensive and time-consuming. The introduction of data concerning the relative binding affinities of one or more DNA-binding molecules to all 256 four base pair DNA sequences allows the development, via molecular modeling, of *ad hoc* protocols for DNA structural analysis and subsequent DNA-binding drug design. The accumulation of data for the DNA sequences to which small molecules bind is likely to result in more accurate, less expensive molecular modeling programs for the analysis of DNA.

The screening capacity of the assay of the present invention is much greater than screening a single DNA sequence with an individual cognate DNA-binding protein. Direct competition assays involving individual receptor:ligand complexes (e.g., a specific DNA:protein complex) are most commonly used for mass screening efforts. Each such assay requires the identification, isolation, purification, and production of the assay components. In particular, a suitable DNA:protein interactions must be identified for each selected screening sequence. Using the assay of the present invention, libraries of synthetic chemicals or biological molecules can be screened to detect molecules that have preferential binding to virtually any specified DNA sequence -- all using a single assay system. When employing the assay of the present invention, secondary screens involving the specific DNA:protein interaction may not be necessary, since inhibitory molecules detected in the assay may be tested directly on a biological system: for example, the ability to disrupt viral replication in a tissue culture or animal model.

6. The Design of New DNA-Binding Heteropolymers Comprised of Subunits Directed to Different DNA Sequences.

The assay of the present invention will facilitate the predictive abilities of molecular modeling systems in two ways. First, *ad hoc* methods of structural prediction will be improved. Second, by employing pattern matching schemes, the comparison of sequences having similar or different affinities for a given set of DNA-binding molecules should empirically reveal sets of sequences that have similar structures (see Section VI.D, Using a Test Matrix). Molecular modeling programs are "trained" using the information concerning DNA-binding molecules and their preferred binding sequences. With this information coupled to the predicative power of molecular modeling programs, the design of DNA-binding molecules (subunits) that could be covalently linked becomes feasible.

These molecular subunits would be directed at defined sections of DNA. For example, a subunit would be designed for each possible DNA unit. For example, if single bases were the binding target of the subunits, then four subunits would be required, one to correspond to each base pair. These subunits could then be linked together to form a DNA-binding polymer, where the DNA binding preference of the polymer corresponds to the sequence binding preferences of the subunits in the particular order in which the subunits are assembled.

Another example of such a polymer is using subunits whose binding was directed at two base sections of DNA. In this case, $4^2 = 16$ subunits would be used, each subunit having a binding affinity for a specific two base pair sequence (e.g., AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT).

If the polymers were to be comprised of subunits targeted to 3 base pair sections of DNA, then $4^3 = 64$ subunits would be prepared. The design of such molecular subunits is dependent upon the establishment of a refined database using empirical data derived by the method of the present invention, as described in Section VI.B.

C. Sequences Targeted by the Assay.

The DNA:protein assay of the present invention has been designed to screen for compounds that bind a full range of DNA sequences that vary in length as well as complexity. Sequence-specific DNA-binding molecules discovered by the assay have potential usefulness as either molecular reagents, therapeutics, or therapeutic precursors. Sequence-specific DNA-binding molecules are potentially powerful therapeutics for essentially any disease or condition that in some way involves DNA. Examples of test sequences for the assay include: a) binding sequences of factors involved in the maintenance or propagation of infectious agents, especially viruses, bacteria, yeast and other fungi, b) sequences causing the inappropriate expression of certain cellular genes, and c) sequences involved in the replication of rapidly growing cells. Furthermore, gene expression or replication need not necessarily be disrupted by blocking the binding of specific proteins. Specific sequences within protein-coding regions of genes (e.g., oncogenes) are equally valid test sequences since the binding of small molecules to these sequences is likely to perturb the transcription and/or replication of the region. Finally, any molecules that bind DNA with some sequence specificity, that is, not just to one particular test sequence, may be still be useful as anti-cancer agents. Several small molecules with some sequence preference are already in use as anticancer therapeutics. Molecules identified by the present assay may be particularly valuable as lead compounds for the development of congeners having either different specificity or different affinity.

One advantage of the present invention is that the assay is capable of screening for binding activity directed against any DNA sequence. Such sequences can be medically significant target sequences scrambled or randomly generated DNA sequences, or well-defined, ordered sets of DNA sequences. Other sets could be used for screening for molecules demonstrating sequence preferential binding (like Doxorubicin) to determine the sequences with highest binding affinity and/or to determine the relative affinities between a large number of different sequences. There is usefulness in taking either approach for detecting and/or designing new therapeutic agents. Section VI.C.3, "Theoretical Considerations for Choosing Target Sequences", outlines the theoretical considerations for choosing DNA target sites in a biological system.

1. Medically Significant Target Sequences.

Few effective viral therapeutics are currently available; yet several potential target sequences for antiviral DNA-binding drugs have been well-characterized. Furthermore, with the accumulation of sequence data on all biological systems, including viral genomes, cellular genomes, pathogen genomes (bacteria, fungi, eukaryotic parasites, etc.), the number of target sites for DNA-binding drugs will increase greatly in the future.

There are numerous methods for identifying medically significant target sequences for DNA-binding drugs, including, but not limited to, the following.

First, medically significant target sequences are found in pathogens of the biological kingdoms, for example in genetic sequences that are key to biochemical pathways or physiological processes. Second, a target is identified, such as (i) a pathogen involved in an infectious disease, or (ii) a biochemical pathway or physiological process of a noninfectious disease, genetic condition, or other biological process. Then specific genes important for the survival of the pathogen or modulation of the endogenous pathway involved in the target system are identified. Third, specific target sequences are identified that affect the expression or activity of a DNA molecule, such as genes or sites involved in replication.

There are numerous pathogens that are potential targets for DNA-binding drugs designed using the methods described in this application. Table I lists a number of potential target pathogens.

Table I: Pathogens	
VIRUSES	
Retroviruses	
Human	HIV I ,II HTLV I, II
Animal	SIV STLV I FELV FIV BLV BIV (Bovine immunodeficiency virus) Lentiviruses Avian reticuloendotheliosis virus
Animal - continued	SIV STLV I FELV FIV BLV BIV (Bovine immunodeficiency virus) Lentiviruses Avian reticuloendotheliosis virus Avian sarcoma and leukosis viruses Caprine arthritis-encephalitis Equine infectious anemia virus Maedi/visna of sheep MMTV (mouse mammary tumor virus) Progressive pneumonia virus of sheep
Herpesviridae	
Human	EBV CMV
	HSV I, II VZV HH6 Cercopithecine Herpes Virus (B Virus) Old world monkeys with infection into humans.
Animal	Bovine Mammillitis virus Equine Herpes virus Equine coital exanthema virus Equine rhinopneumonitis virus Infectious bovine rhinotracheitis virus Marek's disease virus of fowl Turkey herpesvirus
Hepadnaviruses	
Human	HBV/HDV
Animal	Duck Hepatitis Woodchucks Squirrels
Poxviridae	
Human	Orf virus Cow Pox Variola virus Vaccinia Small Pox Pseudocowpox

Poxviridae - continued	
Animal	Bovine papular stomatitis virus Cowpox virus Ectromelia virus (mouse pox) Fibroma viruses of rabbits/squirrels Fowlpox Lumpy skin disease of cattle virus Myxoma Pseudocowpox virus Sheep pox virus Swine pox
Papovaviridae	
Human	BK virus SV-40 JC virus Human Papillomaviruses 1-58 (see list Fields)
Animal	Lymphotropic papovavirus (LPV) Monkey Bovine papillomavirus Shope papillomavirus
Adenoviridae	
Human	Adenoviruses 1-4
Animal	Canine adenoviruses 2
Parvoviridae	
Human	AAV (Adeno Associated Virus) B19 (human)
Animal	FPV (Feline parvovirus) PPV (Porcine parvovirus) ADV (Aleutian disease, mink) Bovine Parvovirus Canine Parvovirus Feline panleukopenia virus Minute virus of mice Mink enteritis virus
BACTERIA	
Streptococcus	
	pneumonia bovis
Group A Streptococci	
	Agents responsible for: Streptococcal pharyngitis Cervical adenitis Otitis media Mastoiditis Peritonsillar abscesses Meningitis Peritonitis
	Pneumonia Acute glomerulonephritis Rheumatic fever Erythema nodosum
Staphylococcus	
	aureus epidermidis saprophyticus cohnii

<i>haemolyticus</i> <i>xylosus</i> <i>warneri</i>
<i>capitis</i> <i>hominis</i> <i>silmulans</i> <i>saccharolyticus</i> <i>auricularis</i> Agents responsible for: <i>Furuncles</i> <i>Carbuncles</i> <i>Osteomyelitis</i> Deep tissue abscesses Wound infections <i>Pneumonia</i> <i>Empyema</i> <i>Pericarditis</i> <i>Endocarditis</i> <i>Meningitis</i> Purulent arthritis Enterotoxin in food poisoning
<i>Branhamella catarrhalis</i>
<i>Neisseria</i> <i>gonorrhoea</i> <i>lactamica</i> <i>sicca</i> <i>subflava</i> <i>mucosa</i>
<i>Neisseria</i> - continued <i>flavescens</i> <i>cinerea</i> <i>elongata</i> <i>canis</i> <i>meningitides</i>
Enteric Bacilli and Similar Gram-Negative Bacteria <i>Escherichia</i> <i>Proteus</i> <i>Klebsiella</i> <i>Pseudomonas aeruginosa</i> <i>Enterobacter</i> <i>Citrobacter</i> <i>Proteus</i>
<i>Providencia</i> <i>Bacteroides</i> <i>Serratia</i> <i>Pseudomonas</i> (not <i>aeruginosa</i>) <i>Acinetobacter</i> <i>Salmonella</i> <i>Shigella</i> <i>Aeromonas</i> <i>Moraxella</i> <i>Edwardsiella</i> <i>Ewingella</i> <i>Hafnia</i> <i>Kluyvera</i> <i>Morganella</i> <i>Plesiomonas</i>
<i>Pseudomonas</i> <i>aeruginosa</i> <i>putida</i> <i>pseudomallei</i> <i>mallei</i>

<i>Haemophilus</i>	<i>ducreyi</i> <i>influenzae</i> <i>parainfluenzae</i>
<i>Bordetella</i>	<i>pertussis</i>
<i>Yersinia</i>	<i>pestis</i> (plague) <i>pseudotuberculosis</i> <i>enterocolitica</i>
<i>Francisella</i>	<i>tularensis</i>
<i>Pasteurella</i>	<i>multocida</i>
<i>Vibrio</i>	<i>cholerae</i> <i>parhaemolyticus</i> <i>fluvialis</i> <i>furnissii</i> <i>mimicus</i>
<i>Brucella</i>	<i>melitensis</i> <i>abortus</i> <i>suis</i> <i>canis</i>
<i>Bartonella</i>	<i>bacilliformis</i>
<i>Gardnerella</i>	<i>vaginalis</i>
<i>Borrelia</i>	<i>recurrentis</i> <i>hermsii</i> <i>duttoni</i> <i>crocidurae</i> <i>burgdorferi</i> (Lyme disease)
<i>Bacillus</i>	<i>anthracis</i> <i>cereus</i> <i>megaterium</i> <i>subtilis</i> <i>sphaericus</i> <i>circulans</i> <i>brevis</i> <i>lentiiformis</i> <i>macerans</i> <i>pumilus</i> <i>thuringiensis</i> <i>larvae</i> <i>lentimorbus</i> <i>popilliae</i>
<i>Streptobacillus</i>	<i>moniliformis</i> (rat bite fever)
<i>Spirillum</i>	<i>minus</i> (rat bite fever)
<i>Rothia</i>	<i>dentocariosa</i>
<i>Kurthia</i>	
<i>Clostridium</i>	<i>botulinum</i> <i>nouyi</i> <i>bifermentans</i>
<i>Clostridium</i> - continued	<i>histolyticum</i> <i>ramosum</i> <i>tetani</i> <i>perfringens</i> <i>novyi</i> <i>septicum</i>
<i>Campylobacter</i>	<i>jejuni</i> <i>fetus</i>

hyintestinalis fennelliae cinaedi
Corynebacterium ulcerans pseudotuberculosis JK diphtheriae
Legionella pneumophila bosemanii micdadie bosnamii feleii many others
Mycobacterium tuberculosis africanum bovis leprae avium complex kansasii fortuitum complex scrofulaceum marinum ulcerans
Actinomyces
Bacteroides fragilidis
Fusobacterium necrophorum nucleatum
Peptostreptococcus
Arachnia
Bifidobacterium
Propionibacterium
Nocardia
Treponema pallidum (syphilis)
Rickettsiae Typhus R. prowazeki (epidemic) R. prowazeki (Brill's disease) R. typhi (endemic) Spotted fever R. rickettsi R. sibiricus R. conorii R. australis R. akari Scrub typhus R. tsutsugamushi Q fever Coxiella burnetii Trench fever Rochalimaea quintana
Chlamydiae C. trachomatis (blindness, pelvic inflammatory disease, LGV)
Mycoplasma pneumoniae Ureaplasma urealyticum
Cardiobacterium hominis
Actinobacillus actinomycetemcomitans

T U R N I N G P A G E S

<i>Kingella</i>
<i>Capnocytophaga</i>
<i>Pasteurella multocida</i>
<i>Leptospira interrogans</i>
<i>Listeria monocytogenes</i>
<i>Erysipelothrix rhusiopathiae</i>
<i>Streptobacillus moniliformis</i>
<i>Calymmatobacterium granulomatis</i>
<i>Bartonella bacilliformis</i>
<i>Francisella tularensis</i>
<i>Salmonella typhi</i>
FUNGAL
<i>Actinomyces</i>
<i>israelii</i>
<i>naeslundii</i>
<i>viscosus</i>
<i>odontolyticus</i>
<i>meyeri</i>
<i>pyogenes</i>
<i>Cryptococcus neoformans</i>
<i>Blastomyces dermatitidis</i>
<i>Histoplasma capsulatum</i>
<i>Coccidioides immitis</i>
<i>Paracoccidioides brasiliensis</i>
<i>Candida</i>
<i>albicans</i>
<i>tropicalis</i>
(Torulopsis) <i>glabrata</i>
<i>parapsilosis</i>
<i>Aspergillus</i>
<i>fumigatus</i>
<i>flavus</i>
<i>niger</i>
<i>terreus</i>
<i>Rhinosporidiosis seeberi</i>
<i>Phycomycetes</i>
<i>Sporothrix schenickii</i>
<i>Mucorales</i>
<i>Entomophthorales</i>
Agents of Chromoblastomycosis
<i>Microsporum</i>
<i>M. audouinii</i> (ring worm)
<i>M. canis</i>
<i>M. gypseum</i>
<i>Trichophyton</i>
<i>T. schoenleinii</i> (favus-ringworm)
<i>T. violaceum</i> (hair)
<i>T. tonsurans</i> (hair)
<i>T. mentagrophytes</i> (athlete's foot)
<i>T. rubrum</i> (athlete's foot)
<i>Malassezia furfur</i>
<i>Cladosporium</i>
<i>werneckii</i>
<i>carrioni</i>
<i>Fonsecaea</i>
<i>pedrosoi</i>
<i>compacta</i>
<i>Phialophora verrucosa</i>
<i>Rhinocladiella aquaspersa</i>
<i>Trichosporon cutaneum</i>
<i>Piedraia hortai</i>

Ascomycota
Basidiomycota
Deuteromycota
Norcardia
brasiliensis
caviae
asteroides
PARASITIC PATHOGENS
Plasmodium (malaria)
falciparum
vivax
ovale
malariae
Schistosoma
japonicum
mansoni
haematobium
intercalatum
mekongi
Trypanosoma
brucei gambiense
brucei rhodesiense
evansi
cruzi
equiperdum
congolense
Entamoeba histolytica
Naegleria fowleri
Acanthamoeba
astronyxis
castellanii
culbertsoni
hatchetti
palestinensis
polyphaga
rhysodes
Leishmania
dovonani
infantum
chagasi
topica
major
aethiopica
mexicana
braziliensis
peruviana
Pneumocystis carinii (interstitial pneumonia)
Babesia (tick born hemoprotozoan)
microti
divergens
Giardia lamblia
Trichomonas (venereal disease)
vaginalis
hominis
tenax
Cryptosporidium parvum (intestinal protozoan)
Isopora belli (dysentery)
Balantidium coli (protozoon induced dysentery)
Dientamoeba fragilis
Blastocystis hominis
Trichinella spiralis (parasitic nematode)
Wuchereria bancrofti (lymphatic filariasis)

<i>Brugia</i> (lymphatic filariasis) <i>malayi</i> <i>timori</i>
<i>Loa loa</i> (eye worm)
<i>Onchocerca volvulus</i>
<i>Mansonella</i> <i>perstans</i> <i>ozzardi</i> <i>streptocerca</i>
<i>Dirofilaria immitis</i>
<i>Angiostrongylus cantonensis</i> <i>costaricensis</i> <i>malayensis</i> <i>mackerrasae</i>
<i>Anisakis</i> (nematode) <i>simplex</i> <i>typica</i>
<i>Pseudoterranova decipiens</i>
<i>Gnathostoma spinigerum</i>
<i>racunculus medinensis</i> (filarial parasite, guinea worm)
<i>Trichuris trichiura</i> (whip worm)
<i>Ascaris lumbricoides</i> (nematode)
<i>Toxocara canis</i> (nematode round worms)
<i>Necator americanus</i> (heart worm)
<i>Ancylostoma</i> (hook worm) <i>duodenale</i> <i>ceylanicum</i> <i>americanus</i>
members of the species <i>Trichostrongylus</i>
<i>Strongyloides</i> (intestinal nematode) <i>stercoralis</i> <i>fuelleborni</i>
<i>Capillaria philippinensis</i> (intestinal nematode)
arious species of <i>Paragonimus</i> (lung fluke disease)
Various species of <i>Micorsporida</i>
<i>Clonorchis sinensis</i> (liver fluke)
<i>Fasciola</i> (trematode, intestinal worm) <i>hepatica</i> <i>gigantica</i>
<i>Fasciolopsis buski</i>
<i>Heterophyes heterophyes</i>
<i>Metagonimus yakagawa</i>
<i>Taenia</i> <i>saginata</i> (beef tapeworm) <i>solium</i> (pork tapeworm)
<i>Hymenolepis</i> (dwarf tapeworm) <i>nana</i> <i>nana fraterna</i> <i>diminuta</i>
<i>Dipylidium caninum</i> (tapeworm of dogs and cats)
<i>Diphyllbothrium</i> (fish tapeworms) <i>lantum</i> <i>dalliae</i> <i>nihonkaiense</i> <i>pacificum</i>
<i>Echinococcus</i> (tape worm with cysts) <i>granulosus</i> <i>multilocularis</i> <i>vogeli</i>
<i>Enterobius vermicularis</i> (Pin worm)

In addition to pathogens, many non-infectious diseases may be controlled at the level of DNA. These diseases are therefore potential candidates for treatment with DNA-binding therapeutics that are discovered or designed using the methods described in this application. Table II lists a number of potential non-infectious diseases that may be targeted for treatment using DNA-binding molecules.

5

Table II: Noninfectious Diseases	
CANCER	
Lung	
Adenocarcinoma	
Squamous cell	
Small cell	
Breast carcinoma	
Ovarian	
Serous tumors	
Mucinous tumors	
Endometrioid carcinoma	
Endometrial carcinoma	
Colon carcinoma	
Malignant Melanoma	
Prostate carcinoma	
Lymphoma	
Hodgkins	
Non-Hodgkin's	
Leukemia	
Chronic Myelogenous	
Acute Myelogenous	
Chronic Lymphocytic	
Acute Lymphocytic	
Cervical carcinoma	
Seminoma	
Multiple Myeloma	
Bladder carcinoma	
Pancreatic carcinoma	
Stomach carcinoma	
Thyroid	
Papillary adenocarcinoma	
Follicular carcinoma	
Medullary carcinoma	
Oral & Pharyngeal carcinomas	
Laryngeal carcinoma	
Bladder carcinoma	
Renal cell carcinoma	
Hepatocellular carcinoma	
Glioblastoma	
Astrocytoma	
Meningioma	
Osteosarcoma	
Pheochromocytoma	
CARDIOVASCULAR DISEASES	
Hypertension	
Essential	
Malignant	
Acute Myocardial Infarction	
Stroke	
Ischemic	
Hemorrhagic	
Angina Pectoris	
Unstable angina	
Congestive Heart Failure	

Supraventricular arrhythmias
Ventricular arrhythmias
Deep Venous Thrombosis
Pulmonary Embolism
Hypercholesterolemia
Cardiomyopathy
Hypertriglyceridemia
RESPIRATORY DISORDERS
Allergic rhinitis
Asthma
Emphysema
Chronic bronchitis
Cystic Fibrosis
Pneumoconiosis
Respiratory distress syndrome
Idiopathic pulmonary fibrosis
Primary pulmonary hypertension
GASTROINTESTINAL DISORDERS
Peptic ulcers
Cholelithiasis
Ulcerative colitis
Crohn's disease
Irritable Bowel Syndrome
Gastritis
Gilbert's syndrome
Nausea
ENDOCRINE/METABOLIC DISORDERS
Diabetes mellitus type I
Diabetes mellitus type II
Diabetes insipidus
Hypothyroidism
Hyperthyroidism
Gout
Wilson's disease
Addison's disease
Cushing's syndrome
Acromegaly
Dwarfism
Prolactinemia
Morbid obesity
Hyperparathyroidism
Hypoparathyroidism
Osteomalacia
RHEUMATOLOGY/IMMUNOLOGY DISORDERS
Transplant rejection
Systemic lupus erythematosus
Rheumatoid arthritis
Temporal Arteritis
Amyloidosis
Sarcoidosis
Sjogren's Syndrome
Scleroderma
Ankylosing spondylitis
Polymyositis
Reiter's Syndrome
Polyarteritis nodosa
Kawasaki's disease

HEMATOLOGIC DISORDERS	
Anemia	
Sickle cell	
Sideroblastic	
Hereditary spherocytosis	
Aplastic	
Autoimmune hemolytic anemia	
Thalassemia	
Disseminated intravascular coagulation	
Polycythemia vera	
Thrombocytopenia	
Thrombotic thrombocytopenic purpura	
Idiopathic thrombocytopenic purpura	
Hemophilia	
von Willebrand's disease	
Neutropenia	
Post-chemotherapy	
Post-radiation	
NEUROLOGIC DISORDERS	
Alzheimer's disease	
Parkinson's disease	
Myasthenia gravis	
Multiple sclerosis	
Amyotrophic lateral sclerosis	
Epilepsy	
Headaches	
Migraine	
Cluster	
Tension	
Guillain-Barre syndrome	
Pain (post-op, trauma)	
Vertigo	
PSYCHIATRIC DISORDERS	
Anxiety	
Depression	
Schizophrenia	
Substance abuse	
Manic-Depression	
Anorexia	
DERMATOLOGIC DISORDERS	
Acne	
Psoriasis	
Eczema	
Contact dermatitis	
Pruritis	
OPHTHALMIC DISORDERS	
Glaucoma	
Allergic conjunctivitis	
Macular degeneration	
MUSCULOSKELETAL DISORDERS	
Osteoporosis	
Muscular dystrophy	
Osteoarthritis	
GENETIC DISORDERS	
Down's syndrome	
Marfan's syndrome	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

Neurofibromatosis
Tay-Sachs disease
Gaucher's disease
Niemann-Pick disease
GENITAL-URINARY DISORDERS
Benign prostatic hypertrophy
Polycystic kidney disease
Non-infectious glomerulonephritis
Goodpasture's syndrome
Urolithiasis
Endometriosis
Impotence
Infertility
Fertility control
Menopause

Once a disease or condition is identified as a potential candidate for treatment with a DNA-binding therapeutic, specific genes or other DNA sequences that are crucial for the expression of the disease associated gene (or survival of a pathogen) are identified within the biochemical or physiological pathway (or the pathogen). In humans, many genes involved in important biological functions have been identified. Virtually any DNA sequence is a potential target site for a DNA-binding molecule, including mRNA coding sequences, promoter sequences, origins of replication, and structural sequences, such as telomeres and centromeres. One class of sites that may be preferable are the recognition sequences for proteins that are involved in the regulation or expression of genetic material. For this reason, the promoter/regulatory regions of genes also provide potential target sites (Table III, see also Example 15).

Table III: Human Genes with Promoter Regions that are Potential Targets for DNA-Binding Molecules

*[LOCUS Names are from EMBL database ver. 33. 1992.]

LOCUS Names*	Locus Description
>HS5FDX	Human ferredoxin gene, 5' end.
>HSA1ATCA	Human macrophage alpha1-antitrypsin cap site region
>HSA1GPB1	Human gene B for alpha 1-acid glycoprotein exon 1 and 5' flank
>HSA1MBG1	Human gene for alpha-1-micro-globulin-bikunin, exons 1-5 (encoding
>HSA2MGLB1	H.sapiens gene for alpha-2 macroglobulin, exon 1
>HSACAA1	H.sapiens ACAA gene (exons 1 & 2) for peroxisomal 3-oxoacyl-CoA
>HSACCOA	Homo sapiens choline acetyltransferase gene sequence.
>HSACEB	Human angiotensin I-converting enzyme (ACE) gene, 5' flank.
>HSACHG1	Human gene fragment for the acetylcholine receptor gamma subunit
>HSACT2CK1	Human cytokine (Act-2) gene, exon 1.
>HSACTBPR	Human beta-actin gene 5'-flanking region
>HSACTCA	Human cardiac actin gene, 5' flank.
>HSACTSA	Human gene for vascular smooth muscle alpha-actin (ACTSA), 5'
>HSACTSG1	Human enteric smooth muscle gamma-actin gene, exon 1.
>HSAD12L	Human arachidonate 12-lipoxygenase gene, 5' end.

>HSADH1X	Human alcohol dehydrogenase alpha subunit (ADH1) gene, exon 1.
>HSADH2X	Human alcohol dehydrogenase beta subunit (ADH2) gene, exon 1.
>HSAFPCP	Human alpha-fetoprotein gene, complete cds.
>HSAK1	Human cytosolic adenylate kinase (AK1) gene, complete cds.
>HSAGAL	Human alpha-N- acetylgalactosaminidase (NAGA) gene, complete cds.
>HSALADG	H.sapiens ALAD gene for porphobilinogen synthase
>HSALBENH	Human albumin gene enhancer region.
>HSALDA1	Human aldolase A gene 5' non-coding exons
>HSALDCG	Human aldolase C gene for fructose-1,6-bisphosphate aldolase
>HSALDOA	Human aldolase A gene (EC 4.1.2.13)
>HSALDOBG	Human DNA for aldolase B transcription start region
>HSALIFA	Human leukemia inhibitory factor (LIF) gene, complete cds.
>HSAMINON	Human aminopeptidase N gene, complete cds.
>HSAMY2A1	Human alpha-amylase (EC 3.2.1.1) gene AMY2A 5-flank and exon 1
>HSAMYB01	Human amyloid-beta protein (APP) gene, exon 1. 1154
>HSANFG1	Human gene fragment for pronatriodilatin precursor (exons 1 and 2)
>HSANFPRE	Human gene for atrial natriuretic factor (hANF) precursor
>HSANFZ1	Human atrial natriuretic factor gene, complete cds.
>HSANGG1	Human angiotensinogen gene 5' region and exon 1
>HSANT1	Human heart/skeletal muscle ATP/ADP translocator (ANT1) gene,
>HSAPC3A	Human apolipoprotein CIII gene and apo AI-apo CIII intergenic
>HSAPC3G	Human gene for apolipoprotein C-III
>HSAPOA2	Human gene for apolipoprotein AII
>HSAPOAIA	Human fetal gene for apolipoprotein AI precursor
>HSAPOBPRM	Human apoB gene 5' regulatory region (apolipoprotein B)
>HSAPOC2G	Human apoC-II gene for preproapolipoprotein C-II
>HSAPOCIA	Human apolipoprotein C-I (VLDL) gene, complete cds.
>HSAPOLIDG	H.sapiens promoter region of gene for apolipoprotein D
>HSARG1	Human arginase gene exon 1 and flanking regions (EC 3.5.3.1)
>HSASG5E	Human argininosuccinate synthetase gene 5' end 1105
>HSATP1A3S	Human sodium/potassium ATPase alpha 3 subunit (ATP1 A3) gene, 5'
>HSBSF2	Human (BSF-2/IL6) gene for B cell stimulatory factor-2
>HSC5GN	Human C5 gene, 5' end. 650
>HSCAII	Human gene fragment for carbonic anhydrase II (exons 1 and 2)
>HSCALCAC	Human calcitonin/alpha-CGRP gene

>HSCALRT1	Human DNA for calretinin exon 1
>HSCAPG	Human cathepsin G gene, complete cds.
>HSCAVIII1	H.sapiens carbonic anhydrase VII (CA VII) gene, exon 1.
>HSCBMYHC	Human gene for cardiac beta myosin heavy chain
>HSCD3AA	Human complement C3 protein mRNA, 5' flank. >HSCD4 Human recognition/surface antigen (CD4) gene, 5' end.
>HSCD44A	Human hyaluronate receptor (CD44) gene, exon 1.
>HSCFTC	Human cystic fibrosis transmembrane conductance regulator gene, 5'
>HSCH7AHYR	Human cholesterol 7-alpha-hydroxylase (CYP7) gene, 5' end.
>HSCHAT	Human gene for choline acetyltransferase (EC 2.3.1.6), partial
>HSCHYMASE	Human mast cell chymase gene, complete cds.
>HSCHYMB	Human heart chymase gene, complete cds. 3279
>HSCKBG	Human gene for creatine kinase B (EC 2.7.3.2)
>HSCNP	Human C-type natriuretic peptide gene, complete cds.
>HSCD59011	Human transmembrane protein (CD59) gene, exon 1.
>HSCDPRO	Human myeloid specific CD11b promoter DNA.
>HSCETP1	Human cholesteryl ester transfer protein. (CETP) gene, exons 1 and
>HSCFTC	Human cystic fibrosis transmembrane conductance regulator gene, 5'
>HSCOSEG	H.sapiens coseg gene for vasopressin-neurophysin precursor
>HSCREKIN	Human creatine kinase gene, exon 1.
>HSCRYABA	Human alpha-B-crystallin gene, 5' end.
>HSCS5P	Human C3 gene, 5' end.
>HSCSF1G1	Human gene for colony stimulating factor CSF-1 5' region
>HSCSPA	Human cytotoxic serine proteinase gene, complete cds.
>HSCST3G	Human CST3 gene for cystatin C
>HSCST4	H.sapiens CST4 gene for Cystatin D
>HSCYP2C8	Human CYP2C8 gene for cytochrome P-450, 5' flank and exon 1
>HSCYP45A	Human gene for cholesterol desmolase cytochrome P-450(SCC) exon 1
>HSCYPB1	Human steroid 11-beta-hydroxylase (CYP11B1) gene, exons 1 and 2.
>HSCYPXI	Human CYPXI gene for steroid 18-hydroxylase (P-450 C18). 2114
>HSCYPXIB1	Human CYPXIB gene for steroid 11beta-hydroxylase (P-450 11beta),
>HSCYPXIX	Human CYPXIX gene, exon 1 coding for aromatase P-450 (EC 1.14.14.1)
>HSDAFC1	Human decay-accelerating factor (DAF) gene, exons 1 and 2.
>HSDBH1	Human DNA for dopamine beta-hydroxylase exon 1 (EC 1.14.17.1)
>HSDDES	Human desmin gene, complete cds.
>HSDKERB	Human cytokeratin 8 (CK8) gene, complete cds.
>HSDNAPOL	Human DNA polymerase alpha gene, 5' end.

>HSDOPAM	H.sapiens dopamine D1A receptor gene, complete exon 1, and exon 2,
>HSECP1	Human DNA for eosinophil cationic protein ECP
>HSEGFA1	Human HER2 gene, promoter region and exon 1.
>HSEL20	Human elastin gene, exon 1.
>HSELAM1B	Human endothelial leukocyte adhesion molecule I (ELAM-1) gene,
>HSEMBPA	Human eosinophil major basic protein gene, complete cds.
>HSENBK1	Human preproenkephalin B gene 5' region and exon 1
>HSENO35	Human ENO3 gene 5' end for muscle-specific enolase
>HSEOSDN	Human DNA for eosinophil derived neurotoxin
>HSEPR	Human erythropoietin receptor mRNA sequence derived from DNA, 5'
>HSERB2P	Human c-erb B2/neu protein gene, 5'end, and promoter region.
>HSERCC25	Human genomic and mRNA sequence for ERCC2 gene 5'region involved in
>HSERPA	Human erythropoietin gene, complete cds.
>HSERR	Human mRNA for oestrogen receptor
>HSESTEI1	H.sapiens exon 1 for elastase I
>HSFBRGG	Human gene for fibrinogen gamma chain
>HSFCERG5	Human lymphocyte IgE receptor gene 5'-region (Fc-epsilon R)
>HSFERG1	Human apoferritin H gene exon 1
>HSFIBBR1	Human fibrinogen beta gene 5' region and exon 1
>HSFIXG	Human factor IX gene, complete cds.
>HSFKBP1	Human FK506 binding proteins 12A, 12B and 12C (FKBP12) mRNA, exons
>HSFLAP1	Human 5-lipoxygenase activating protein (FLAP) gene, exon 1.
>HSFOS	Human fos proto-oncogene (c-fos), complete cds.
>HSGOS2PE	Human GOS2 gene, upstream region and cds.
>HSGCSFG	Human gene for granulocyte colony-stimulating factor (G-CSF)
>HSGEGR2	Human EGR2 gene 5' region 1233
>HSGHPROM	Human growth hormone (hGH) gene promoter
>HSGIPX1	Human gastric inhibitory polypeptide (GIP) mRNA, exon 1.
>HSGLA	Human GLA gene for alpha-D-galactosidase A (EC 3.2.1.22)
>HSGLUC1	Human glucagon gene transcription start region 732
>HSGMCSFG	Human gene for granulocyte-macrophage colony stimulating factor
>HSGR1	Human glucocorticoid receptor gene, exon 1. 1602
>HSGRFP1	Human growth hormone-releasing factor (GRF) gene, exon 1 (complete)
>HSGSTP15	Human GST pi gene for glutathione S-transferase pi exon 1 to 5
>HSGTRH	Human gene for gonadotropin-releasing hormone
>HSGYPC	Human glycoporphin C (GPC) gene, exon 1, and promoter region.
>HSH10	Human histone (H10) gene, 5' flank.
>HSH1DNA	Human gene for H1 RNA 1057
>HSH1FNC1	Human H1 histone gene FNC16 promoter region

>HSH2B2H2	Human H2B.2 and H2A.1 genes for Histone H2A and H2B
>HSH4AHIS	H.sapiens H4/a gene for H4 histone
>HSH4BHIS	H.sapiens H4/b gene for H4 histone
>HSHARA	Human androgen receptor gene, transcription initiation sites.
>HSHCG5B1	Human chorionic gonadotropin (hCG) beta subunit gene 5 5'-flank
>HSEMPRO	Human DNA for hemopoxin promoter
>HSHIAPPA	Human islet amyloid polypeptide (hIAPP) gene, complete cds.
>HSHIH4	Human H4 histone gene
>HSHISH2A	Human histone H2a gene
>HSHISH2B	Human histone H2b gene
>HSHISH3	Human histone H3 gene
>HSHLAA1	Human HLA-A1 gene
>HSHLAB27	Human gene for HLA-B27 antigen
>HSHLABW	Human HLA-Bw57 gene
>HSHLAF	Human HLA-F gene for human leukocyte antigen F
>HSHLIA	Human gene for histocompatibility antigen HLA-A3
>HSHLIC	Human gene for class I histocompatibility antigen HLA-CW3
>HSHMG17G	Human HMG-17 gene for non-histone chromosomal protein HMG-17
>HSHOX3D	Human HOX3D gene for homeoprotein HOX3D
>HSHSC70	Human hsc70 gene for 71 kd heat shock cognate protein
>HSHSP70D	Human heat shock protein (hsp 70) gene, complete cds.
>HSHSP70P	Human hsp70B gene 5'-region
>HSHIAPP12	Human IAPP gene exon 1 and exon 2 for islet amyloid polypeptide
>HSICAMAB	Human intercellular adhesion molecule 1 (ICAM-1) gene, exon 1.
>HSIFI54	Human interferon-inducible gene IFI-54K 5'flank
>HSIFNA14	Human interferon alpha gene IFN-alpha 14
>HSIFNA16	Human interferon alpha gene IFN-alpha 16
>HSIFNA5	Human interferon alpha gene IFN-alpha 5
>HSIFNA6	Human interferon alpha gene IFN-alpha 6
>HSIFNA7	Human interferon alpha gene IFN-alpha 7
>HSIFNG	Human immune interferon (IFN-gamma) gene.
>HSIFNIN6	Human alpha/beta-interferon(IFN)-inducible 6-16 gene exon 1 and
>HSIGF24B	Human DNA for insulin-like growth factor II (IGF-2); exon 4B
>HSIGFBP1A	Human insulin-like growth factor binding protein (hIGFBP1) gene
>HSIGK10	Human germline gene for the leader peptide and variable region
>HSIGK15	Human germline gene for the leader peptide and variable region
>HSIGK17	Human rearranged gene for kappa immunoglobulin subgroup V kappa IV
>HSIGK20	Human rearranged DNA for kappa immunoglobulin subgroup V kappa III
>HSIGKLC1	Human germline fragment for immunoglobulin kappa light chain
>HSIGVA5	Human germline immunoglobulin kappa light

	chain V-segment
>HSIL05	Human interleukin-2 (IL-2) gene and 5'-flanking region
>HSIL1AG	Human gene for interleukin 1 alpha (IL-1 alpha)
>HSIL1B	Human gene for prointerleukin 1 beta
>HSIL2RG1	Human interleukin 2 receptor gene 5' flanking region and exon 1
>HSIL45	Human interleukin 4 gene 5'-region
>HSIL5	Human interleukin 5 (IL-5) gene, complete cds.
>HSIL6B	Human interleukin 6 (IL 6) gene, 5' flank.
>HSIL71	Human interleukin 7 (IL7) gene, exon 1.
>HSIL9A	Human IL9 protein gene, complete cds.
>HSINSU	Human gene for preproinsulin, from chromosome 11. Includes a highly
>HSINT1G	Human int-1 mammary oncogene
>HSJUNCAA	Human jun-B gene, complete cds.
>HSKER65A	Human DNA for 65 kD keratin type II exon 1 and 5' flank
>HSKERUHS	Human gene for ultra high-sulphur keratin protein
>HSLACTG	Human alpha-lactalbumin gene
>HSLAG1G	Human LAG-1 gene
>HSLCATG	Human gene for lecithin-cholesterol acyltransferase (LCAT)
>HSLCK1	Human lymphocyte-specific protein tyrosine kinase (lck) gene
>HSLFACD	Human leukocyte function-associated antigen-1 (LFA-1 or CD11a)
>HSLPLA	Human lipoprotein lipase (LPL) gene, 5' flank.
>HSLYAM01	Human leukocyte adhesion molecule-1 (LAM-1), exon 1.
>HSLYSOZY	Human lysozyme gene (EC 3.2.1.17)
>HSMBP1A	Human DNA for mannose binding protein 1 (MBP1), Exon 1
>HSMCCPAA	Human mast cell carboxypeptidase A (MC-CPA) gene, exons 1-2.
>HSMDR1	Human P-glycoprotein (MDR1) mRNA, complete cds.
>HSMED	Human bone marrow serine protease gene (medullasin)
>HSMEHG	Human DNA (exon 1) for microsomal epoxide hydrolase
>HSMETIE	Human metallothionein-Ie gene (hMT-Ie).
>HSMG01	Human myoglobin gene (exon 1)
>HSMGSAG	Human gene for melanoma growth stimulatory activity (MGSA)
>HSMHCAG1	Human alpha-MHC gene for myosin heavy chain N-terminus)
>HSMHCGE1	Human class II invariant gamma-chain gene (5' flank, exon 1)
>HSMHCW5	Human MHC class I HLA-Cw5 gene, 5' flank.
>HSMNLN1	Human motilin gene exon 1
>HSMPOA	Human myeloperoxidase gene, exons 1-4.
>HSMRP	Human mitochondrial RNA-processing endoribonuclease RNA (mrp) gene
>HSMTS1A	H.sapiens mts1 gene, 5' end.
>HSMYCE12	Human myc-oncogene exon 1 and exon 2
>HNAKATP	Human Na,K-ATPase beta subunit (ATP1B) gene,

	exons 1 and 2.
>HSNEURK1	H.sapiens gene for neuromedin K receptor (exon 1)
>HSNFH1	Human gene for heavy neurofilament subunit (NF-H) exon 1
>HSNFIL6	Human gene for nuclear factor NF-IL6
>HSNFLG	Human gene for neurofilament subunit NF-L
>HSNK21	Human neurokinin-2 receptor (NK-2) gene, exon 1.
>HSNMYC	Human germ line n-myc gene
>HSNRASPR	H. sapiens N-RAS promoter region
>HSODC1A	Human ornithine decarboxylase (ODC1) gene, complete cds.
>HSOTCEX1	Human ornithine transcarbamylase (OTC) gene, 5'-end region.
>HSOTNPI	Human prepro-oxytocin-neurophysin I gene, complete cds.
>HSP450SCC	Human cytochrome P450scc gene, 5' end and promoter region.
>HSP53G	Human p53 gene for transformation related protein p53
>HSPADP	Human promoter DNA for Alzheimer's disease amyloid A4 precursor
>HSPAI11	Human gene for plasminogen activator inhibitor 1 (PAI-1) 5'-flank
>HSPGDF	Human platelet-derived growth factor A-chain (PDGF) gene, 5' end
>HSPGP95G	Human PGP9.5 gene for neuron-specific ubiquitin C-terminal
>HSPLSM	Human plasminogen gene, exon 1.
>HSPNMTB	Human gene for phenylethanolamine N-methylase (PNMT) (EC 2.1.1.28)
>HSPOMC5F	Human opiomelanocortin gene, 5' flank.
>HSPPI4B	Human placental protein 14 (PP14) gene, complete cds.
>HSPRB3L	Human gene PRB3L for proline-rich protein G1
>HSPRB4S	Human PRB4 gene for proline-rich protein Po, allele S
>HSPRLNC	Human prolactin mRNA, partial cds.
>HSPROAA1	Human prothymosin-alpha gene, complete cds.
>HSPROT2	Human protamine 2 gene, complete cds.
>HSPRPE1	Human SPR2-1 gene for small proline rich protein (exon 1)
>HSPS2G1	Human estrogen-responsive gene pS2 5'flank and exon 1
>HSPSAP	Human pulmonary surfactant apoprotein (PSAP) gene, complete cds.
>HSPSP94A	Human gene for prostatic secretory protein PSP-94, exon 1
>HSPTHRPA	Human parathyroid hormone-related peptide (PTHRP) gene, exons 1A,
>HSPURNPHO	Human gene for purine nucleoside phosphorylase (upstream region)
>HSRDNA	Human rDNA origin of transcription
>HSREGA01	Human regenerating protein (reg) gene, complete cds.
>HSREN01	Human renin gene 5' region and exon 1
>HSRPBG1	Human gene fragment for retinol binding protein (RBP) (exon 1-4)
>HSSAA1A	Human serum amyloid A (GSAA1) gene, complete cds.

>HSSAA1B	H.sapiens SAA1 beta gene
>HSSB4B1	Human gene fragment for HLA class II SB 4-beta chain (exon 1)
>HSSI5G5	Human c-sis proto-oncogene 5' region
>HSSLIPG	Human SLPI gene for secretory leukocyte protease inhibitor
>HSSOD1G1	Human superoxide dismutase (SOD-1) gene exon 1 and 5' flanking
>HSSODB	Human ornithine decarboxylase gene, complete cds.
>HSSRDA01	H.sapiens steroid 5-alpha-reductase gene, exon 1.
>HSSUBP1G	H.sapiens gene for substance P receptor (exon 1)
>HSSYB1A1	Human synaptobrevin 1 (SYB1) gene, exon 1.
>HSTAT1	Human gene for tyrosine aminotransferase (TAT) (EC 2.6.1.5) Exon 1.
>HSTCBV81	Human T-cell receptor V-beta 8.1 gene 775
>HSTCRB21	Human T-cell receptor beta chain gene variable region.
>HSTFG5	Human transferrin (Tf) gene 5' region
>HSIL3FL5	Human interleukin 3 gene, 5' flank.
>HSTFPB	Human tissue factor gene, complete cds.
>HSTGFB1	Human mRNA for transforming growth factor-beta (TGF-beta)
>HSTGFB3B	Human transforming growth factor beta-3 gene, 5' end.
>HSTGFBET2	Human transforming growth factor beta-2 gene, 5' end.
>HSTH01	Human tyrosine hydroxylase (TH) (EC 1.14.16.2) gene from upstream
>HSTHIO2A	Human metallothionein gene IIA promoter region
>HSTHRO01	Human thrombospondin gene, exons 1, 2 and 3.
>HSTHXBG	H.sapiens gene for thyroxine-binding globulin gene
>HSTHYR5	Human thyroglobulin gene 5' region
>HSTNFA	Human gene for tumor necrosis factor (TNF-alpha)
>HSTNFB	Human gene for lymphotoxin (TNF-beta)
>HSTOP01	Homo sapiens type I DNA topoisomerase gene, exons 1 and 2.
>HSTPIA	Human triosephosphate isomerase (TPI) gene, 5' end.
>HSTPO5	Human thyroid peroxidase gene 5' end (EC 1.11.1.7)
>HSTRP	Human transferrin receptor gene promoter region
>HSTRPY1B	Human tryptase-I gene, complete cds.
>HSTUBB2	Human beta 2 gene for beta-tubulin
>HSTYRO1E	Human tyrosinase gene, exon 1 and 5' flanking region (EC 1.14.18.1)
>HSU6RNA	Human gene for U 6 RNA
>HSUPA	Human uPA gene for urokinase-plasminogen activator
>HSVAVP01	Human proto-oncogene vav, 5' end.
>HSVCAM1A	Human vascular cell adhesion molecule-1 (VCAM1) gene, complete CDS.
>HSVIM5RR	Human vimentin gene 5' regulatory region

Once the gene target or, in the case of small pathogens, the genome target has been identified, short sequences within the gene or genome target are identified as medically significant target sites. Medically significant target sites can be defined as short DNA sequences (approximately 4-30 base pairs) that are required for the expression or replication of genetic material. For example, sequences that bind regulatory factors, either transcriptional or replicatory factors, are ideal target sites for altering gene or viral expression.

Further, coding sequences may be adequate target sites for disrupting gene function, although the disruption of a polymerase complex that is moving along the DNA sequence may require a stronger binder than for the disruption of the initial binding of a regulatory protein.

Finally, even non-coding, non-regulatory sequences may be of interest as target sites (e.g., for disrupting replication processes or introducing an increased mutational frequency).

Several specific examples of medically significant target sites are shown in Table IV.

Table IV
MEDICALLY SIGNIFICANT DNA-BINDING SEQUENCES

Test sequence	DNA-binding Protein	Medical Significance
EBV origin of replication	EBNA	Infectious mononucleosis, nasal pharyngeal carcinoma
HSV origin of replication	UL9	Oral and genital Herpes
VZV origin of replication	UL9-like	Shingles
HPV origin of replication	E2	Genital warts, cervical carcinoma
Interleukin 2 enhancer	NFAT-1	Immunosuppressant
HIV LTR	NFAT-1 NFkB	AIDS, ARC
HBV enhancer	HNF-1	Hepatitis
Fibrogen promoter	HNF-1	Cardiovascular disease
Oncogene promoter and coding sequences	??	cancer

(Abbreviations: EBV, Epstein-Barr virus; EBNA, Epstein-Barr virus nuclear antigen; HSV, Herpes Simplex virus; VZV, Varicella zoster virus; HPV, human papilloma virus; HIV LTR, Human immunodeficiency virus long terminal repeat; NFAT, nuclear factor of activated T cells; NFkB, nuclear factor kappaB; AIDS acquired immune deficiency syndrome; ARC, AIDS related complex; HBV, hepatitis B virus; HNF, hepatic nuclear factor.)

For example, origin of replication binding proteins have short, well-defined binding sites within the viral genome and are therefore excellent target sites for a competitive DNA-binding drug. Examples of such proteins include, Epstein Barr virus nuclear antigen 1 (EBNA-1) (Ambinder, et al.; Reisman, et al.), E2 (which is encoded by the human papilloma virus) (Chin, et al.), UL9 (which is encoded by herpes simplex virus type 1) (McGeoch, et al.), and the homologous protein in varicella zoster virus (VZV) (Stow, et al.).

Similarly, recognition sequences for DNA-binding proteins that act as transcriptional regulatory factors are also good target sites for antiviral DNA-binding drugs. Examples listed in Table IV include (i) the binding site for hepatic nuclear factor (HNF-1), which is required for the expression of human hepatitis B virus (HBV) (Chang), and (ii) NFkB and NFAT-1 binding sites in the human immunodeficiency virus (HIV) long terminal repeat (LTR), one or both of which may be involved in the expression of the virus (Greene, W.C.).

Examples of non-viral DNA targets for DNA-binding drugs are also shown in Table IV to illustrate the wide range of potential applications for sequence-specific DNA-binding molecules. For example, nuclear factor of activated T cells (NFAT-1) is a regulatory factor that is crucial to the inducible expression of the interleukin 2 gene in response to signals from the antigen receptor, which, in turn, is required for the cascade of molecular events during T cell activation (for review, see Edwards, C.A., and Crabtree, G.R.). The mechanism of action of two immunosuppressants, cyclosporin A and FK506, is thought to be to block the inducible expression of NFAT-1 (Schmidt, et al. and Banerji, et al.). However, the effects of these drugs are not specific to NFAT-1; therefore, a drug targeted specifically to the NFAT-1 binding site in the IL-2 enhancer would be desirable as an improved immunosuppressant.

Targeting the DNA site with a DNA-binding drug rather than targeting with a drug that affects the DNA-binding protein (presumably the target of the current immunosuppressants) is advantageous for at least two reasons: first, there are many fewer target sites for specific DNA sequences than specific proteins (e.g., in the case of glucocorticoid receptor, a handful of DNA-binding sites vs. about 50,000 protein molecules in each cell); and second, only the targeted gene need be affected by a DNA-binding drug, while a protein-binding drug would disable all the cellular functions of the protein. An example of the latter point is the binding site for HNF-1 in the human fibrinogen promoter. Fibrinogen level is one of the most highly correlated factor with cardiovascular disease. A drug targeted to either HNF-1 or the HNF-1 binding site in the fibrinogen promoter might be used to decrease fibrinogen expression in individuals at high risk for disease because of the over-expression of fibrinogen. However, since HNF-1 is required for the expression of a number of normal hepatic genes, blocking the HNF-1 protein would be toxic to liver function. In contrast, by blocking a DNA sequence that is composed in part of the HNF-1 binding site and in part by flanking sequences for divergence, the fibrinogen gene can be targeted with a high level of selectivity, without harm to normal cellular HNF-1 functions.

The assay has been designed to screen virtually any DNA sequence. Test sequences of medical significance include viral or microbial pathogen genomic sequences and sequences within or regulating the expression of oncogenes or other inappropriately expressed cellular genes. In addition to the detection of potential antiviral drugs, the assay of the present invention is also applicable to the detection of potential drugs for (i) disrupting the metabolism of other infectious agents, (ii) blocking or reducing the transcription of inappropriately expressed cellular genes (such as oncogenes or genes associated with certain genetic disorders), and (iii) the enhancement or alteration of expression of certain cellular genes.

2. Defined Sets of Test Sequences.

The approach described in the above section emphasizes screening large numbers of fermentation broths, extracts, or other mixtures of unknowns against specific medically significant DNA target sequences. The assay can also be utilized to screen a large number of DNA sequences against known DNA-binding drugs to determine the relative affinity of the single drug for every possible defined specific sequence. For example, there are 4^n possible sequences, where n = the number of nucleotides in the sequence. Thus, there are $4^3 = 64$ different three base pair sequences, $4^4 = 256$ different four base pair sequences, $4^5 = 1024$ different 5 base pair sequences, etc. If these sequences are placed in the test site, the site adjacent to the screening sequence (the example used in this invention is the UL9 binding site), then each of the different test sequences can be screened against many different DNA-binding molecules.

The test sequences may be placed on either or both sides of the screening sequence, and the sequences flanking the other side of the test sequences are fixed sequences to stabilize the duplex and, on the 3' end of the top strand, to act as an annealing site for the primer (see Example 1). In Figure 14B, the TEST and SCREENING sequences are indicated. The preparation of such double-stranded oligonucleotides is described in Example 1 and illustrated in Figure 4.

The test sequences, denoted in Figure 14B as X:Y (where X = A,C,G, or T and Y = the complementary sequence, T,G,C, or A), may be any of the 256 different 4 base pair sequences shown in Figure 13.

Once a set of test oligonucleotides containing all possible four base pair sequences has been synthesized (see Example 1), the set can be screened with any DNA-binding drug. The relative effect of the drug on each oligonucleotide assay system will reflect the relative affinity of the drug for the test sequence. The entire spectrum of affinities for each particular DNA sequence can therefore be defined for any particular DNA-binding drug. This data, generated using the assay of the present invention, can be used to facilitate molecular modeling programs and/or be used directly to design new DNA-binding molecules with increased affinity and specificity.

Another type of ordered set of oligonucleotides that may be useful for screening are sets comprised of scrambled sequences with fixed base composition. For example, if the recognition sequence for a protein is 5'-GATC-3' and libraries were to be screened for DNA-binding molecules that recognized this sequence, then it would be desirable to screen sequences of similar size and base composition as control sequences for the assay. The most precise experiment is one in which all possible 4 bp sequences are screened. In the case of a 4 base-pair sequence, this represents $4^4 = 256$ different test sequences: a number of screening sequences that may not be practical in every situation. However, there are many fewer possible 4 bp sequences with the same base composition (1G, 1A, 1T, 1C) ($n!$) = 24 different 4 bp sequences with this particular base composition), such sequences provide excellent controls without having to screen large numbers of sequences.

3. Theoretical Considerations in Choosing Biological Target Sites: Specificity and Toxicity.

One consideration in choosing sequences to screen using the assay of the present invention is test sequence accessibility, that is, the potential exposure of the sequence *in vivo* to binding molecules. Cellular DNA is packaged in chromatin, rendering most sequences relatively inaccessible. Sequences that are actively transcribed, particularly those sequences that are regulatory in nature, are less protected and more accessible to both proteins and small molecules. This observation is substantiated by a large literature on DNAase I sensitivity, footprinting studies with nucleases and small molecules, and general studies on chromatin structure (Tullius). The relative accessibility of a regulatory sequence, as determined by DNAase I hypersensitivity, is likely to be several orders of magnitude greater than an inactive portion of the cellular genome. For this reason the regulatory sequences of cellular genes, as well as viral regulatory or replication sequences, are useful regions to choose for selecting specific inhibitory small molecules using the assay of the present invention.

Another consideration in choosing sequences to be screened using the assay of the present invention is the uniqueness of the potential test sequence. As discussed above for the nuclear protein HNF-1, it is desirable that small inhibitory molecules are specific to their target with minimal cross reactivity. Both sequence composition and length effect sequence uniqueness. Further, certain sequences are found less frequently in the human genome than in the genomes of other organisms, for example, mammalian viruses. Because of base composition and codon utilization differences, viral sequences are distinctly different from mammalian sequences. As one example, the dinucleotide CG is found much less frequently in mammalian cells than the dinucleotide sequence GC: further, in SV40, a mammalian virus, the sequences AGCT and ACGT are represented 34 and 0 times, respectively. Specific viral regulatory sequences can be chosen as test sequences keeping this bias in mind. Small inhibitory molecules identified which bind to such test sequences will be less likely to interfere with cellular functions.

There are approximately 3×10^9 base pairs (bp) in the human genome. Of the known DNA-binding drugs for which there is crystallographic data, most bind 2-5 bp sequences. There are $4^4 = 256$ different 4 base sequences; therefore, on average, a single 4 bp site is found roughly 1.2×10^7 times in the human genome. An individual 8 base site would be found, on average, about 50,000 times in the genome. On the surface, it might appear that drugs targeted at even an 8 bp site might be deleterious to the cell because there are so many binding sites; however, several other considerations must be recognized.

First, most DNA is tightly wrapped in chromosomal proteins and is relatively inaccessible to incoming DNA-binding molecules as demonstrated by the nonspecific endonucleolytic digestion of chromatin in the nucleus (Edwards, C.A. and Firtel, R.A.). Active transcription units are more accessible, but the most highly exposed regions of DNA in chromatin are the sites that bind regulatory factors. As demonstrated by DNAase I hypersensitivity (Gross, D.S. and Garrard, W.T.), regulatory sites may be 100-1000 times more sensitive to endonucleolytic attack than the bulk of chromatin. This is one reason for targeting regulatory sequences with DNA-binding drugs.

Secondly, several anticancer drugs that bind 2, 3, or 4 bp sequences have sufficiently low toxicity so that they can be used as drugs. This indicates that, if high affinity binding sites for known drugs can be matched with specific viral target sequences, it may be possible to use currently available drugs as antiviral agents at lower concentrations than they are currently used, with a concomitantly lower toxicity.

4. Further Considerations in Choosing Target Sites: Finding Eukaryotic Promoters.

Eukaryotic organisms have three RNA polymerases (Pol I, II, and III) that transcribe genetic information from DNA to RNA. The correct regulation of this information flow is essential for the survival of the cell. These multi-subunit enzymes need additional proteins to regulate transcription. Many of these additional proteins bind to DNA in a region 5' of the translation start site for a gene: this region is generally known as the promoter region of the gene.

All three polymerases use a core set of general transcription proteins to bind to this region. A central component of this complex is the protein called TBP or TFIID. The site this protein binds to is known as the TATA-box because the sequence usually contains a sequence motif similar to TATA (e.g., TATAA/tAa/t). Originally it was thought that each of the three polymerases used a separate set of general transcription factors and that Pol II used TFIID exclusively. Recently it has been shown that all three classes of RNA polymerase need TFIID for transcriptional regulation (see Comai, et al.; and Greenblatt)

A molecule that binds to a DNA sequence closely adjacent or overlapping a TATA binding site will likely alter transcriptional regulation of the gene. If the molecule binds based solely on specificity to the TATA-box sequence itself, then this molecule is expected to be very toxic to cells since the transcription of most genes would be altered. The sequences adjacent to TATA boxes, however, are not conserved. Accordingly, if a particular sequence is selected adjacent a TATA box of a particular gene, a molecule that binds to

this specific sequence would be expected to alter the transcriptional regulation of just that gene.

TATA-boxes were first identified by determining the sequence of the DNA located 5' of the RNA start sites of a number of genes. Examination of these sequences revealed that most genes had a TATA-box motif (consensus sequence) in the range of nucleotides 50 to 15 nucleotides 5' of the RNA start site. In vitro studies, typically DNA protection (footprinting) studies, lead to the conclusion that proteins were binding to these sites. Further in vitro DNA binding experiments demonstrated that some proteins could specifically bind to these sites. This lead to assays that allowed purification and subsequent sequencing of the binding proteins. This information facilitated the cloning and expression of genes encoding the binding proteins. A large number of transcription factors are now known. The protein designated TFIID has been demonstrated to bind to the TATA-box (Lee, et al.).

Molecules that interfere with the interaction of these transcription factors and their target DNA (i.e., DNA/Protein transcription complexes) are also expected to alter transcription initiated from the target DNA. A publicly available database of these factors and the sequences to which they bind is available from the National Library of Medicine and is called "The Transcription Data Base, or TFD." The binding sites of these transcription factors can be identified in the 5' non-coding region of genes having known sequences (Example 15).

The ability to select target sequences adjacent the binding site of a transcription factor, as described above for TFIID, can be applied to other general transcription factors as well. For the purpose of the present invention, a general transcription factor is one that regulates the transcriptional expression of more than one gene. For any such general transcription factor, as for TFIID above, a particular target sequence can be selected adjacent the transcription factor binding site of a selected gene. A molecule that binds to this specific target sequence would be expected to alter the transcriptional regulation of just that gene and not all of the genes for which the transcription factor regulates expression. Alteration of transcriptional regulation may involve inhibition or increased affinity (enhancement) of binding of a transcription factor to its cognate DNA.

Many examples of such general transcription factors have been identified, including, but not limited to, the following: SP1 (Raney, et al., 1992; Kitadai, et al., 1992); NFAT-1 (Shaw, et al., 1988); Ets family of transcription factors, including E1f1 (Thompson, et al., 1992); Fos protein (Neuberg, et al., 1991); NF-kappa (Wirth, et al., 1988; Meijer, et al., 1992); and AP1-like proteins, including the product of the c-jun oncogene (Descheemaeker, et al., 1992; Ryder et al., 1988; Harshman et al., 1988; Angel et al., 1988; Bos et al., 1988; Bohmann et al., 1987).

Accordingly, for a selected gene, non-conserved DNA surrounding the transcription factor binding site can be chosen as a specific target sequence for small molecule binding. A small molecule can be chosen whose binding overlaps an adjacent transcription factor DNA binding sequence (e.g., by 1-3 nucleotide pairs). In this case, the specificity of DNA binding for the small molecule is, in large part, derived from the non-conserved sequences adjacent the transcription factor binding site, in order to reduce small molecule binding at the transcription factor binding site associated with other genes.

Small molecules that bind such specific target sequences can be identified and/or designed using the assay and methods of the present invention.

5. Further Considerations in Choosing Alternative Small-Molecule-Binding Target Sites.

Small molecules that interfere with the interaction of any DNA binding protein and its cognate DNA (i.e., DNA/Protein complexes) can be selected by the assay and methods of the present invention. As described above for general transcription factors, sequences adjacent the DNA binding site for a selected DNA binding protein can serve as a target for small molecule binding in order to alter the interaction of the DNA binding protein and its cognate site. The small molecule can affect the DNA:protein interaction, for example, by inhibiting or enhancing the association of protein with the DNA.

For a selected DNA:protein interaction, non-conserved DNA surrounding the selected DNA binding site can be chosen as a specific target sequence for small molecule binding. In some cases the small molecule binding can overlap the DNA binding site: for example, in the case of a therapeutic used to treat a mammal with a bacterial infection, a small molecule may be selected to bind to the bacterial origin of DNA replication. Such a small molecule may essentially completely overlap the region defined by the bacterial origin-of-replication-DNA:protein interaction since a corresponding target sequence is not likely present in the DNA of the mammalian host.

However, in the case where selective binding is required, as described above for TFIID, the specificity of the small molecule for DNA binding should essentially derive from the non-conserved sequences adjacent the DNA-binding protein's cognate DNA-binding site. This results in small molecule binding being reduced at similar DNA:protein binding sites at other locations.

6. Further Considerations in Choosing Target Sites: Procaryotes and Viruses.

Bacterial gene expression is regulated at several different levels, including transcription. General and specific transcription factors are needed along with the core RNA polymerase to accurately produce appropriate amounts of mRNA. Antibiotics that bind to the RNA polymerase and prevent mRNA production are potent bacterial poisons: molecules that could interfere with the initiation of transcription for specific essential genes are expected to have similar effects.

Many bacterial promoters have been sequenced and carefully examined. In general, the majority of bacterial promoters have two well characterized regions, the -35 region which has a consensus sequence similar to SEQ ID NO:625 and the -10 region with a consensus sequence of SEQ ID NO:626. The sequence of the start site for RNA polymerase, however, is not always the same. The start site is determined by a supplementary protein called the sigma factor, which confers specificity for binding the RNA polymerase core. Several sigma factors are present in any species of bacteria. Each sigma factor recognizes a different set of promoter sequences. Expression of sigma factors is regulated, typically, by the growth conditions the bacteria is encountering. These sigma factor promoter sequences represent excellent targets for sequence specific DNA binding molecules.

As an example of choosing target sequences for the purpose of designing a DNA-binding therapeutic for a bacterial disease, consider the example of tuberculosis. Tuberculosis is caused by *Mycobacterium tuberculosis*.

All bacteria need to make ribosomes for the purpose of protein synthesis. The -35 and -10 regions of *M. tuberculosis* ribosome RNA synthesis has been determined. In the EMBL locus MTRRNOP the -35 signal is located at coordinants 394..400 and the -10 signal is found at coordinants 419..422. These regions represent excellent targets for a DNA binding drug that would inhibit the growth of the bacteria by disrupting its ability to make ribosomes and synthesize protein. Multiple other essential genes could be targeted in a similar manner.

M. tuberculosis is a serious public health problem for several reasons, including the development of antibiotic resistant strains. Many antibiotics inhibit the growth of bacteria by binding to a specific protein and inhibiting its function. An example of this is the binding of rifampicin to the beta subunit of the bacterial RNA polymerase. Continued selection of bacteria with an agent of this kind can lead to the selection of mutants having an altered RNA polymerase so that the antibiotic can no longer bind it. Such mutants can arise from a single mutation.

However, binding a drug to a DNA regulatory region requires at least two mutations to escape the inhibitory effect of the drug: one mutation in the target DNA sequence so that the drug could not bind the target sequence, and one mutation in the regulatory binding protein so that it can recognize the new, mutated regulatory sequence. Such a double mutation event is much less frequent than the single mutation discussed above, for example, with rifampicin. Accordingly, it is expected that the development of drug resistant bacteria would be much less common for DNA-binding drugs that bind to promoter sequences.

The HIV viral promoter region (shown in Figure 28) provides an example of choosing DNA target sequences for sequence-specific DNA binding drugs to inhibit viral replication.

Many eukaryotic viruses use promoter regions that have similar features to normal cellular genes. The replication of these viruses depends on the general transcription factors present in the host cell. As such, the promoter sequences in DNA viruses are similar to those found in cellular genes and have been well-studied. The binding factors Sp-1 and TFIID are important generalized factors that most viral promoters use.

In the HIV promoter sequence found in LOCUS HIVBH101 in version 32 of the EMBL databank, three tandem decanucleotide Sp1 binding sites are located between positions 377 and 409. Site III shows the strongest affinity for the cellular factor. The three cause up to a tenfold effect on transcriptional efficiency in vitro. The transcription start site is at position 455, with a TATA box at 427-431 in the sequence listed below. In addition to these sites, there are two NF-kappa-B sites in this region between nucleotides 350 and 373. These sites are annotated in Figure 28.

Sequence-specific DNA binding molecules that specifically disrupted this binding would be expected to disrupt HIV replication. For example, the sequences adjacent to the TFIID binding site (SEQ ID NO:628 and/or SEQ ID NO:629), would be target sites for a DNA-binding molecule designed to disrupt TFIID binding. These sequences are found in HIV but are not likely to occur overlapping TFIID binding sites in the endogenous human genome. Multiple sites could be targeted to decrease the likelihood that a single mutation could prevent drug binding.

D. Using Test Matrices and Pattern Matching for the Analysis of Data.

The assay described herein has been designed to use a single DNA:protein interaction to screen for sequence-specific and sequence-preferential DNA-binding molecules that can recognize virtually any specified sequence. By using sequences flanking the recognition site for a single DNA:protein interaction, a very large number of different sequences can be tested. The analysis of data yielded by such experiments displayed as matrices and analyzed by pattern matching techniques should yield information about the relatedness of DNA sequences.

The basic principle behind the DNA:protein assay of the present invention is that when molecules bind DNA sequences flanking the recognition sequence for a specific protein the binding of that protein is blocked. Interference with protein binding likely occurs by either (or both) of two mechanisms: (i) directly by steric hindrance, or (ii) indirectly by perturbations transmitted to the recognition sequence through the DNA molecule.

Both of these mechanisms will presumably exhibit distance effects. For inhibition by direct steric hindrance direct data for very small molecules is available from methylation and ethylation interference studies. These data suggest that for methyl and ethyl moieties, the steric effect is limited by distance effects to 4-5 base pairs. Even still the number of different sequences that can theoretically be tested for these very small molecules is still very large (i.e., 5 base pair combinations total 4^5 (=1024) different sequences).

In practice, the size of sequences tested can be explored empirically for different sized test DNA-binding molecules. A wide array of sequences with increasing sequence complexity can be routinely investigated. This may be accomplished efficiently by synthesizing degenerate oligonucleotides and multiplexing oligonucleotides in the assay process (i.e., using a group of different oligonucleotides in a single assay) or by employing pooled sequences in test matrices.

In view of the above, assays employing a specific protein and oligonucleotides containing the specific recognition site for that protein flanked by different sequences on either side of the recognition site can be used to simultaneously screen for many different molecules, including small molecules, that have binding preferences for individual sequences or families of related sequences. Figure 12 demonstrates how the analysis of a test matrix yields information about the nature of competitor sequence specificity. As an example, to screen for molecules that could preferentially recognize each of the 256 possible tetranucleotide sequences (Figure 13), oligonucleotides could

be constructed that contain these 256 sequences immediately adjacent to a 11 bp recognition sequence of UL9 oriS SEQ ID NO:615, which is identical in each construct.

- 5 In Figure 12 "+" indicates that the mixture retards or blocks the formation of DNA:protein complexes in solution and "-" indicates that the mixture had no marked effect on DNA:protein interactions. The results of this test are shown in Table V.

Table V

Test Mix	Specificity
#1,4,7: oligos	none detected for the above
#2: for recognition site	either nonspecific or specific
#3	AGCT
#5	CATT or ATT
#6	GCATTC, GCATT, CATTC, GCAT, or ATT
#8	CTTT

10 These results demonstrate how such a matrix provides data on the presence of sequence specific binding activity is a test mixture and also provides inherent controls for non-specific binding. For example, the effect of test mix #8 on the different test assays reveals that the test mix preferentially affects the oligonucleotides that contain the sequence CCCT. Note that the sequence does not have to be within the test site for test mix #8 to exert an affect. By displaying the data in a matrix, the analysis of the sequences affected by the different test mixtures is facilitated.

15 Furthermore, defined, ordered sets of oligonucleotides can be screened with a chosen DNA-binding molecule. The results of these binding assays can then be examined using pattern matching techniques to determine the subsets of sequences that bind the molecule with similar binding characteristics. If the structural and biophysical properties (such as, geometric shape and electrostatic properties) of sequences are similar, then it is likely that they will bind the molecule with similar binding characteristics. If the structural and biophysical properties of sequences are different, then it is likely that they will not bind the molecule with similar binding characteristics. In this context, the assay might be used to group defined, ordered sequences into subsets based on their binding characteristics: for example, the subsets could be defined as high affinity binding sites, moderate affinity binding sites, and low affinity binding sites. Sequences in the subsets with positive attributes (e.g., high affinity binding) have a high probability of having similar structural and biophysical properties to one another.

20 By screening and analyzing the binding characteristics of a number of DNA-binding molecules against the same defined set of DNA sequences, data can be accumulated about the subsets of sequences that fall into the same or similar subsets. Using this pattern matching approach, which can be computer-assisted, the sequences with similar structural and biophysical properties can be grouped empirically.

25 The database arising from pattern matching analysis of raw assay data will lead to the increased understanding of sequence structure and thereby lead to the design of novel DNA-binding molecules with related but different binding activities.

45 E. Applications for the Determination of the Sequence Specificity of DNA-Binding Drugs.

Applications for the determination of the sequence specificity of DNA-binding drugs are described below. The applications are divided into drug homo- and heteromeric polymers (part 1) and sequence-specific DNA-binding molecules as facilitators of triple strand formation (part 2).

50 One utility of the assay of the invention is the identification of highest affinity binding sites among all possible sites of a certain length for a given DNA-binding molecule. This information may be valuable to the design of new DNA-binding therapeutics.

1. Multimerization of Sequence-Preferential or Sequence-Specific DNA-Binding Molecules Identified in the Assay.

Any particular DNA-binding small molecule screened in the assay may only recognize a 2-4 base pair site, and even if the recognition is quite specific, the molecule may be toxic because there are so many target sites in the genome ($3 \times 10^9/4^4$ 4 bp sites, for example). However, if drugs with differential affinity for different sites are identified, the toxicity of DNA-binding drugs may be drastically reduced by creating dimers, trimers, or multimers with these drugs (Example 13). From theoretical considerations of the free energy changes accompanying the binding of drugs to DNA, the intrinsic binding constant of a dimer should be the square of the binding constant of the monomer (Le Pecq, J.B.). Experimental data confirmed this expectation in 1978 with dimer analogs of ethidium bromide (Kuhlmann, et al.). Dimerization of several intercalating molecules, in fact, yields compounds with DNA affinities raised from 10^5 M⁻¹ for the corresponding monomer to 10^8 to 10^9 M⁻¹ for the dimers (Skorobogaty, et al.; Gaugain, et al. (1978a and b); Le Pecq, et al.; Pelaprat, et al.). Trimerization, which theoretically should yield binding affinities that are the cube of the affinity of the homomonomeric subunit or the product of affinities of the heteromonomeric subunits, has yielded compounds with affinities as high as 10^{12} M⁻¹ (Laugaa, et al.). Such affinity is markedly better than the affinities seen for many DNA regulatory proteins.

As a hypothetical example, if a relatively weak DNA-binding drug, drug X, which binds a 4 bp site with an affinity of 2×10^5 M⁻¹ was dimerized, the bis-X drug would now recognize an 8 bp site with a theoretical affinity of 4×10^{10} M⁻¹. The difference in affinity between the monomer X and the bis-X form is 200,000-fold. The number of 4 bp sites in the genome is approximately 1.2×10^7 versus the number of 8 bp sites in the genome which is approximately 5×10^4 . Accordingly, there are 256-fold fewer 8 bp sites than 4 bp sites. Thus, the number of high affinity target sites is 256-fold fewer for the bis-X molecule than the number of low affinity target sites for the monomer X, with a 200,000-fold difference in affinity between the two types of sites.

Since the binding constant of a dimer is the product of the binding constants of the monomers, when monomers with higher initial binding constants are formed into dimers (or multimers) the differential effect is proportionately increased, creating a wider "window" of affinity versus the number of binding sites. The breadth of the window essentially reflects the margin of effective drug concentration compared to the relative toxicity.

There are two immediate ramifications of dimerization (or multimerization) of monomeric drugs with moderate toxicity and sequence preference. First, the concentration of drug needed is lowered because of the higher affinity, so that even relatively toxic molecules can be used as drugs.

Second, since toxicity is likely linked to the average number of drug molecules bound to the genome, as specificity is increased by increasing the length of the binding site, toxicity is decreased.

Given the information already available on sequence-preferential binding of DNA-binding drugs, it is likely that each drug presented to the screening assay will have (i) a number of high affinity binding sites (e.g., 10 to 100-fold better affinity than the average site), (ii) a larger number of sites that are bound with moderate affinity (3 to 10-fold better affinity than average), (iii) the bulk of the binding sites having average affinity, and (iv) a number of sites having worse-than-average affinity. This range of binding affinities will likely resemble a bell-shaped curve. The shape of the curve will probably vary for each drug. To exemplify, assume that approximately five 4 bp sites will be high affinity binding sites, and twenty 4 bp sites will be moderately high affinity binding sites, then any given drug may recognize roughly 25, high or moderately high affinity binding sites. If 50 to 100 drugs are screened, this represents a "bank" of potentially 250-500 high affinity sites and 1000-2500 moderately high affinity sites. Thus, the probability of finding a number of high affinity drug binding sites that match medically significant target sites is good. Furthermore, heterodimeric drugs can be designed to match DNA target sites of 8 or more bp, lending specificity to the potential pharmaceuticals.

As discussed above, once the sequence preferences are known, the information may be used to design oligomeric molecules (homopolymers or heteropolymers) with substantially greater sequence specificity and substantially higher binding affinity. For example, if a DNA-binding molecule, X, binds a 4 bp sequence 5'-ACGT-3'/5'-ACGT-3' with an equilibrium affinity constant of $2 \times 10^5 \text{ M}^{-1}$, then the dimer of X, X_2 , should bind the dimer of the sequence, 5'-ACGTACGT-3'/5'-ACGTACGT-3', with an equilibrium affinity constant of $(2 \times 10^5 \text{ M}^{-1})^2 = 4 \times 10^{10} \text{ M}^{-2}$. The DNA-binding dimer molecule, X_2 , recognizes an 8 bp sequence, conferring higher sequence specificity, with a binding affinity that is theoretically 200,000-fold higher than the DNA-binding monomer, X.

The same argument can be extended to trimer molecules: the trimer of X, X_3 , would bind a 12 bp sequence, 5'-ACGTACGTACGT-3'/5'-ACGTACGTACGT-3', with a theoretical equilibrium affinity constant of $8 \times 10^{15} \text{ M}^{-2}$.

DNA-binding polymers constructed using the above-mentioned approach may be homo- or hetero-polymers of the parent compounds or oligomeric compounds composed of mixed subunits of the parent compounds. Homopolymers are molecules constructed using two or more subunits of the same monomeric DNA-binding molecule. Heteropolymers are molecules constructed using two or more subunits of different monomeric DNA-binding molecules. Oligomeric compounds are constructed of mixed pieces of parent compounds and may be hetero- or homomeric.

For example, distamycin is a member of a family of non-intercalating minor groove DNA-binding oligopeptides that are composed of repeating units of N-methylpyrrole groups. Distamycin has 3 N-methylpyrrole groups. Examples of homopolymers would be bis-distamycin, the dimer of distamycin, a molecule containing 6 N-methylpyrrole groups or tris-distamycin, the trimer of distamycin, a molecule containing 9 N-methylpyrrole groups.

Daunomycin is a member of an entirely different class of DNA-binding molecules, the anthracycline antibiotics, that bind to DNA via intercalation.

Heteropolymers are molecules composed of different types of DNA-binding subunits; for example, compounds composed of a distamycin molecule linked to a daunomycin molecule or a distamycin molecule linked to two daunomycin molecules. The term "oligomeric" is being used to describe molecules comprised of linked subunits each of which may be smaller than the parent compound.

An example of an homo-oligomeric compound would be a distamycin molecule linked to 1 or 2 additional N-methylpyrrole groups; the resulting molecule would not be as large as bis-distamycin, but would fundamentally be composed of the same component organic moieties that comprise the parent molecule. Examples of a hetero-oligomeric compounds would be daunomycin linked to one or two N-methylpyrrole groups.

The construction of these polymers will be directed by the information derived from the sequence preferences of the parent compounds tested in the assay. In one embodiment of the assay, a database of preferred sequences is constructed, providing a source of information about the 4 bp sequences that bind with relatively higher affinity to particular drugs that may be linked together to target any particular larger DNA sequence.

DNA-binding subunits can be chemically coupled to form heteropolymers or homopolymers. The subunits can be joined directly to each other, as in the family of distamycin molecules, or the subunits can be joined with a spacer molecule, such as carbon chains or peptide bonds. The coupling of subunits is dependent on the chemical nature of the subunits: appropriate coupling reactions can be determined for any two subunit molecules from the chemical literature. The choice of subunits will be directed by the sequence to be targeted and the data accumulated through the methods discussed in Section VI.B of this application.

2. Sequence-Specific DNA-Binding Molecules Identified in the Assay as Facilitators of Triplex Formation.

Several types of nucleic acid base-containing polymers have been described that will form complexes with nucleic acids (for reviews, see Helene, C. and Toulme, J.-J.). One type of such a polymer forms a triple-stranded complex by the insertion of a third strand into the major groove of the DNA helix. Several types of base-recognition specific interactions of third strand

oligonucleotide-type polymers have been observed. One type of specificity is due to Hoogsteen bonding (Hoogsteen). This specificity arises from recognition between pyrimidine oligonucleotides and double-stranded DNA by pairing thymine and adenine:thymine base pairs and protonated cytosine and guanine:cytosine base pairs (Griffin, et al.). Another type of specific interaction involves the use of purine oligonucleotides for triplex formation. In these triplexes, adenine pairs with adenine:thymine base pairs and guanine with guanine:cytosine (Cooney, et al.; Beal and Dervan) or thymine:adenine base pairs (Griffin, L., and Dervan, P.B.).

Other motifs for triplex formation have been described, including the incorporation of nucleic acid analogs (eg, methylphosphonates, phosphorothioates; Miller, et al.), and the invention of backbones other than the phosphoribose backbones normally found in nucleic acids (Pitha, et al.; Summerton, et al.). In several cases, the formation of triplex has been demonstrated to inhibit the binding of a DNA-binding protein (e.g., Young, et al.; Maher, et al.) or the expression of a cellular protein (Cooney, et al.).

Furthermore, several experiments have been reported in which a small DNA-binding molecule has been covalently attached to polymer capable of forming a triplex structure: (i) an acridine:polypyrimidine molecule has been demonstrated to inhibit SV40 in CV-1 cells (Birg, et al.); (ii) cleavage at a single site in a yeast chromosome was achieved with an oligonucleotide:EDTA-Fe molecule (Strobel, et al.; Dervan); and (iii) a photoinducible endonuclease was created by similar strategy by attaching an ellipticine derivative to a homopyrimidine oligonucleotide (Perouault, et al.). Several other small intercalating agents coupled to oligonucleotides have been described (for review, see Montenay-Garestier).

One utility of the assay of the present invention is to identify the sequence-specificity of DNA-binding molecules for use in designing and synthesizing heteromeric therapeutics consisting of a DNA-binding polymer (e.g., an oligonucleotide) attached to a sequence-preferential or sequence-specific DNA-binding molecule, yielding a heteropolymer. The attached small molecule may serve several functions.

First, if the molecule has increased affinity for a specific site (such as, a particular 4 base pair sequence) over all other sites of the same size, then the local concentration of the hetero-molecule, including the oligonucleotide, will be increased at those sites. The amount of heteropolymer, containing a sequence-specific moiety attached to one end, needed for treatment purposes is reduced compared to a heteropolymer that has a non-specific DNA-binding moiety attached. This reduction in treatment amount is directly proportional to both the differential specificity and the relative affinities between the sequence-specific binder and the non-specific binder. For the simplest example, if a sequence-specific molecule with absolute specificity (i.e., it binds only one sequence) had equal affinity for a specific 4 base-pair target site (1/256 possible combinations) as a non-specific molecule, then the amount of drug needed to exert the same effective concentration at that site could potentially be as much as 256-fold less for the specific and non-specific drugs. Accordingly, attaching a sequence-specific DNA-binding molecule to a polymer designed to form triplex structures allows increased localized concentrations.

A second utility of the assay of the present invention is to identify small molecules that cause conformational changes in the DNA when they bind. The formation of triplex DNA requires a shift from B form to A form DNA. This is not energetically favorable, necessitating the use of increased amounts of polymer for triplex formation to drive the conformational change. However, the insertion of a small DNA-binding molecule (such as, actinomycin D), which induces a conformational change in the DNA, reduces the amount of polymer needed to stabilize triplex formation.

Accordingly, one embodiment of the invention is to use the assay to test known DNA-binding molecules with all 256 possible four base pair test sequences to determine the relative binding affinity to all possible 4 bp sequences. Then, once the sequence preferences are known, the information may be used to design heteropolymeric molecules comprised of a small DNA-binding molecule and a macromolecule, such as a triplex-forming oligonucleotide, to obtain a DNA-binding molecule with enhanced binding characteristics. The potential

advantages of attaching a sequence-specific or sequence-preferential DNA-binding small molecule to a triplex forming molecule are to (i) target the triplex to a subset of specific DNA sequences and thereby (ii) anchor the triplex molecule in the vicinity of its target sequence and in doing so, (iii) increase the localized concentration of the triplex molecule, which allows (iv) lower concentrations of triplex to be used effectively. The presence of the small molecule may also facilitate localized perturbations in DNA structure, such as destabilizing the B form of DNA, which is unsuitable for triplex formation. Such destabilization may facilitate the formation of other structures, such a form DNA useful for triplex formation. The net effect would be to decrease the amount of triplex needed for efficacious results.

F. Other Applications.

The potential pharmaceutical applications for sequence-specific DNA-binding molecules are very broad, including antiviral, antifungal, antibacterial, antitumor agents, immunosuppressants, and cardiovascular drugs. Sequence-specific DNA-binding molecules can also be useful as molecular reagents as, for example, specific sequence probes.

As more DNA-binding molecules are detected, information about their DNA binding affinities, sequence recognition, and mechanisms of DNA-binding will be gathered, eventually facilitating the design and/or modification of new molecules with different or specialized activities.

Although the assay has been described in terms of the detection of sequence-specific DNA-binding molecules, the reverse assay could be achieved by adding DNA in excess to protein to look for peptide sequence specific protein-binding inhibitors.

The following examples illustrate, but in no way are intended to limit, the present invention.

Materials and Methods

Synthetic oligonucleotides were prepared using commercially available automated oligonucleotide synthesizers. Alternatively, custom designed synthetic oligonucleotides may be purchased, for example, from Synthetic Genetics (San Diego, CA). Complementary strands were annealed to generate double-strand oligonucleotides.

Restriction enzymes were obtained from Boehringer Mannheim (Indianapolis IN) or New England Biolabs (Beverly MA) and were used as per the manufacturer's directions.

Distamycin A and Doxorubicin were obtained from Sigma (St. Louis, MO). Actinomycin D was obtained from Boehringer Mannheim or Sigma.

Standard cloning and molecular biology techniques are described in Ausubel, et al., and Sambrook, et al.

Example 1

Preparation of the Oligonucleotide Containing the Screening Sequence

This example describes the preparation of (A) biotinylated/digoxigenin/radiolabeled, and (B) radiolabeled double-stranded oligonucleotides that contain the screening sequence and selected Test sequences.

A. Biotinylation.

The oligonucleotides were prepared as described above. The wild-type control sequence for the UL9 binding site, as obtained from HSV, is shown in Figure 4. The screening sequence, i.e. the UL9 binding sequence, is CGTTCGCACTT (SEQ ID NO:601) and is underlined in Figure 4. Typically, sequences 5' and/or 3' to the screening sequence were replaced by a selected Test sequence (Figure 5).

One example of the preparation of a site-specifically biotinylated oligonucleotide is outlined in Figure 4. An oligonucleotide primer complementary to the 3' sequences of the screening sequence-containing oligonucleotide was synthesized. This oligonucleotide terminated at the residue corresponding to the C in position 9 of the screening sequence. The primer oligonucleotide was hybridized to the oligonucleotide containing the screening sequence. Biotin-11-dUTP (Bethesda Research Laboratories (BRL), Gaithersburg MD) and Klenow

enzyme were added to this complex (Figure 4) and the resulting partially double-stranded biotinylated complexes were separated from the unincorporated nucleotides using either pre-prepared "G-25 SEPHADEX" spin columns (Pharmacia, Piscataway NJ) or "NENSORB" columns (New England Nuclear) as per manufacturer's instructions. The remaining single-strand region was converted to double-strands using DNA polymerase I Klenow fragment and dNTPs resulting in a fully double-stranded oligonucleotide. A second "G-25 SEPHADEX" column was used to purify the double-stranded oligonucleotide. Oligonucleotides were diluted or resuspended in 10 mM Tris-HCl, pH 7.5, 50 mM NaCl, and 1 mM EDTA and stored at -20°C. For radiolabelling the complexes, ^{32}P -alpha-dCTP (New England Nuclear, Wilmington, DE) replaced dCTP for the double-strand completion step.

Alternatively, the top strand, the primer, or the fully double-stranded oligonucleotide have been radiolabeled with γ - ^{32}P -ATP and polynucleotide kinase (NEB, Beverly, MA). Most of our preliminary studies have employed radiolabeled, double-stranded oligonucleotides. The oligonucleotides are prepared by radiolabeling the primer with T4 polynucleotide kinase and γ - ^{32}P -ATP, annealing the "top" strand full length oligonucleotide, and "filling-in" with Klenow fragment and deoxynucleotide triphosphates. After phosphorylation and second strand synthesis, oligonucleotides are separated from buffer and unincorporated triphosphates using "G-25 SEPHADEX" preformed spin columns (IBI, New Haven, CT or Biorad, Richmond CA). This process is outlined in Figure 4. The reaction conditions for all of the above Klenow reactions were as follows: 10 mM Tris-HCl, pH 7.5, 10 mM MgCl_2 , 50 mM NaCl, 1 mM dithioerythritol, 0.33-100 μM deoxytriphosphates, 2 units Klenow enzyme (Boehringer-Mannheim, Indianapolis IN). The Klenow reactions were incubated at 25°C for 15 minutes to 1 hour. The polynucleotide kinase reactions were incubated at 37°C for 30 minutes to 1 hour.

B. End-Labeling with Digoxigenin.

The biotinylated, radiolabelled oligonucleotides or radiolabeled oligonucleotides were isolated as above and resuspended in 0.2 M potassium cacodylate (pH=7.2), 4 mM MgCl_2 , 1 mM 2-mercaptoethanol, and 0.5 mg/ml bovine serum albumin. To this reaction mixture digoxigenin-11-dUTP (an analog of dTTP, 2'-deoxy-uridine-5'-triphosphate, coupled to digoxigenin via an 11-atom spacer arm, Boehringer Mannheim, Indianapolis IN) and terminal deoxynucleotidyl transferase (GIBCO BRL, Gaithersburg, MD) were added. The number of Dig-11-dUTP moieties incorporated using this method appeared to be less than 5 (probably only 1 or 2) as judged by electrophoretic mobility on polyacrylamide gels of the treated fragment as compared to oligonucleotides of known length.

The biotinylated or non-biotinylated, digoxigenin-containing, radiolabelled oligonucleotides were isolated as above and resuspended in 10 mM Tris-HCl, 1 mM EDTA, 50 mM NaCl, pH 7.5 for use in the binding assays.

The above procedure can also be used to biotinylate the other strand by using an oligonucleotide containing the screening sequence complementary to the one shown in Figure 4 and a primer complementary to the 3' end of that molecule. To accomplish the biotinylation Biotin-7-dATP was substituted for Biotin-11-dUTP. Biotinylation was also accomplished by chemical synthetic methods: for example, an activated nucleotide is incorporated into the oligonucleotide and the active group is subsequently reacted with NHS-LC-Biotin (Pierce). Other biotin derivatives can also be used.

C. Radiolabelling the Oligonucleotides.

Generally, oligonucleotides were radiolabelled with gamma- ^{32}P -ATP or alpha- ^{32}P -deoxynucleotide triphosphates and T4 polynucleotide kinase or the Klenow fragment of DNA polymerase, respectively. Labelling reactions were performed in the buffers and by the methods recommended by the manufacturers (New England Biolabs, Beverly MA; Bethesda Research Laboratories, Gaithersburg MD; or Boehringer/Mannheim, Indianapolis IN). Oligonucleotides were separated from buffer and unincorporated triphosphates using "G-25 SEPHADEX" preformed spin columns (IBI, New Haven, CT; or Biorad, Richmond, CA) or "NENSORB" preformed columns (New England Nuclear, Wilmington, DE) as per the manufacturers instructions.

There are several reasons to enzymatically synthesize the second strand.

The two main reasons are that by using an excess of primer, second strand synthesis can be driven to near completion so that nearly all top strands are annealed to bottom strands, which prevents the top strand single strands from folding back and creating additional and unrelated double-stranded structures, and secondly, since all of the oligonucleotides are primed with a common primer, the primer can bear the end-label so that all of the oligonucleotides will be labeled to exactly the same specific activity.

Example 2

Preparation of the UL9 Protein

A. Cloning of the UL9 Protein-Coding Sequences into pAC373.

To express full length UL9 protein a baculovirus expression system has been used. The sequence of the UL9 coding region of Herpes Simplex Virus has been disclosed by McGeoch *et al.* and is available as an EMBL nucleic acid sequence. The recombinant baculovirus AcNPV/UL9A, which contained the UL9 protein-coding sequence, was obtained from Mark Challberg (National Institutes of Health, Bethesda MD). The construction of this vector has been previously described (Olivo, *et al.* (1988, 1989)). Briefly, the *NarI/EcoRV* fragment was derived from pMC160 (Wu, *et al.*). Blunt-ends were generated on this fragment by using all four dNTPs and the Klenow fragment of DNA polymerase I (Boehringer Mannheim, Indianapolis IN) to fill in the terminal overhangs. The resulting fragment was blunt-end ligated into the unique *BamHI* site of the baculoviral vector pAC3T3 (Summers, *et al.*).

B. Cloning of the UL9 Sequence in pVL1393.

The UL9 protein-coding region was cloned into a second baculovirus vector, pVL1393 (Luckow, *et al.*). The 3077 bp *NarI/EcoRV* fragment containing the UL9 gene was excised from vector pEcoD (obtained from Dr. Bing Lan Rong, Eye Research Institute, Boston, MA): the plasmid pEcoD contains a 16.2 kb *EcoRI* fragment derived from HSV-I that bears the UL9 gene (Goldin, *et al.*). Blunt-ends were generated on the UL9-containing fragment as described above. *EcoRI* linkers (10 mer) were blunt-end ligated (Ausubel, *et al.*; Sambrook, *et al.*) to the blunt-ended *NarI/EcoRV* fragment.

The vector pVL1393 (Luckow, *et al.*) was digested with *EcoRI* and the linearized vector isolated. This vector contains 35 nucleotides of the 5' end of the coding region of the polyhedron gene upstream of the polylinker cloning site. The polyhedron gene ATG has been mutated to ATT to prevent translational initiation in recombinant clones that do not contain a coding sequence with a functional ATG. The *EcoRI/UL9* fragment was ligated into the linearized vector, the ligation mixture transformed into *E. coli* and ampicillin resistant clones selected. Plasmids recovered from the clones were analyzed by restriction digestion and plasmids carrying the insert with the amino terminal UL9 protein-coding sequences oriented to the 5' end of the polyhedron gene were selected. This plasmid was designated pVL1393/UL9 (Figure 7).

pVL1393/UL9 was cotransfected with wild-type baculoviral DNA (AcMNPV; Summers, *et al.*) into Sf9 (*Spodoptera frugiperda*) cells (Summers, *et al.*). Recombinant baculovirus-infected Sf9 cells were identified and clonally purified (Summers, *et al.*).

C. Expression of the UL9 Protein.

Clonal isolates of recombinant baculovirus infected Sf9 cells were grown in Grace's medium as described by Summers, *et al.* The cells were scraped from tissue culture plates and collected by centrifugation (2,000 rpm, for 5 minutes, 4°C). The cells were then washed once with phosphate buffered saline (PBS) (Maniatis, *et al.*). Cell pellets were frozen at -70°C. For lysis the cells were resuspended in 1.5 volumes 20 mM HEPES, pH 7.5, 10% glycerol, 1.7 M NaCl, 0.5 mM EDTA, 1 mM dithiothreitol (DTT), and 0.5 mM phenyl methyl sulfonyl fluoride (PMSF). Cell lysates were cleared by ultracentrifugation (Beckman table top ultracentrifuge, TLS 55 rotor, 34 krpm, 1 hr, 4°C). The supernatant was dialyzed overnight at 4°C against 2 liters dialysis buffer (20 mM HEPES, pH 7.5, 10% glycerol, 50 mM NaCl, 0.5 mM EDTA, 1 mM dtt, and 0.1 mM PMSF).

These partially purified extracts were prepared and used in DNA:protein binding experiments. If necessary extracts were concentrated using a "CENTRICON 30" filtration device (Amicon, Danvers MA).

D. Cloning the Truncated UL9 Protein.

The sequence encoding the C-terminal third of UL9 and the 3' flanking sequences, an approximately 1.2 kb fragment, was subcloned into the bacterial expression vector, pGEX-2T (Figure 6). The pGEX-2T is a modification of the pGEX-1 vector of Smith, et al. which involved the insertion of a thrombin cleavage sequence in-frame with the glutathione-S-transferase protein (gst).

A 1,194 bp *Bam*HI/*Eco*RV fragment of pEcoD was isolated that contained a 951 bp region encoding the C-terminal 317 amino acids of UL9 and 243 bp of the 3' untranslated region.

This *Bam*HI/*Eco*RV UL9 carboxy-terminal (UL9-COOH) containing fragment was blunt-ended and *Eco*RI linkers added as described above. The *Eco*RI linkers were designed to allow in-frame fusion of the UL9 protein-coding sequence to the gst-thrombin coding sequences. The linked fragment was isolated and digested with *Eco*RI. The pGEX-2T vector was digested with *Eco*RI, treated with Calf Intestinal Alkaline Phosphatase (CIP) and the linear vector isolated. The *Eco*RI linked UL9-COOH fragment was ligated to the linear vector (Figure 6). The ligation mixture was transformed into *E. coli* and ampicillin resistant colonies were selected. Plasmids were isolated from the ampicillin resistant colonies and analyzed by restriction enzyme digestion. A plasmid which generated a gst/thrombin/UL9-COOH in frame fusion was identified (Figure 6) and designated pGEX-2T/UL9-COOH.

E. Expression of the Truncated UL9 Protein.

E. coli strain JM109 was transformed with pGEX-2T/C-UL9-COOH and was grown at 37°C to saturation density overnight. The overnight culture was diluted 1:10 with LB medium containing ampicillin and grown from one hour at 30°C. IPTG (isopropylthio- β -galactoside) (GIBCO-BRL) was added to a final concentration of 0.1 mM and the incubation was continued for 2-5 hours. Bacterial cells containing the plasmid were subjected to the temperature shift and IPTG conditions, which induced transcription from the *tac* promoter.

Cells were harvested by centrifugation and resuspended in 1/100 culture volume of MTPBS (150 mM NaCl, 16 mM Na₂HPO₄, 4 mM NaH₂PO₄). Cells were lysed by sonication and lysates cleared of cellular debris by centrifugation.

The fusion protein was purified over a glutathione agarose affinity column as described in detail by Smith, et al. The fusion protein was eluted from the affinity column with reduced glutathione, dialyzed against UL9 dialysis buffer (20 mM HEPES pH 7.5, 50 mM NaCl, 0.5 mM EDTA, 1 mM DTT, 0.1 mM PMSF) and cleaved with thrombin (2 ng/ug of fusion protein).

An aliquot of the supernatant obtained from IPTG-induced cultures of pGEX-2T/C-UL9-COOH-containing cells and an aliquot of the affinity-purified, thrombin-cleaved protein were analyzed by SDS-polyacrylamide gel electrophoresis. The result of this analysis is shown in Figure 8. The 63 kilodalton GST/C-UL9 fusion protein is the largest band in the lane marked GST-UL9 (lane 2). The first lane contains protein size standards. The UL9-COOH protein band (lane GST-UL9 + Thrombin, Figure 8, lane 3) is the band located between 30 and 46 kD: the glutathione transferase protein is located just below the 30 kD size standard. In a separate experiment a similar analysis was performed using the uninduced culture: it showed no protein corresponding in size to the fusion protein.

Extracts are dialyzed before use. Also, if necessary, the extracts can be concentrated typically by filtration using a "CENTRICON 30" filter.

Example 3 Binding Assays

A. Band Shift Gels.

DNA:protein binding reactions containing both labelled complexes and free DNA were separated electrophoretically on 4-10% polyacrylamide/Tris-Borate-EDTA (TBE) gels (Fried, et al.; Garner, et al.). The gels were then fixed, dried, and exposed to X-ray film. The autoradiograms of the gels were examined for band shift patterns.

B. Filter Binding Assays.

A second method used particularly in determining the half-lives for oligonucleotide:protein complexes is filter binding (Woodbury, et al.). Nitrocellulose disks (Schleicher and Schuell, BA85 filters) that have been

soaked in binding buffer (see below) were placed on a vacuum filter apparatus.

DNA:protein binding reactions (see below; typically 15-30 μ l) are diluted to 0.5 ml with binding buffer (this dilutes the concentration of components without dissociating complexes) and applied to the discs with vacuum applied. Under low salt conditions the DNA:protein complex sticks to the filter while free DNA passes through. The discs are placed in scintillation counting fluid (New England Nuclear), and the cpm determined using a scintillation counter.

This technique has been adapted to 96-well and 72-slot nitrocellulose filtration plates (Schleicher and Schuell) using the above protocol except (i) the reaction dilution and wash volumes are reduced, and (ii) the flow rate through the filter is controlled by adjusting the vacuum pressure. This method greatly facilitates the number of assay samples that can be analyzed. Using radioactive oligonucleotides, the samples are applied to nitrocellulose filters, the filters are exposed to x-ray film, then analyzed using a Molecular Dynamics scanning densitometer. This system transfers data directly into analytical software programs (e.g., Excel) for analysis and graphic display.

Example 4

Functional UL9 Binding Assay

A. Functional DNA-Binding Activity Assay.

Purified protein was tested for functional activity using band-shift assays. Radiolabelled oligonucleotides (prepared as in Example 1B) that contain the 11 bp recognition sequence were mixed with the UL9 protein in binding buffer (optimized reaction conditions: 0.1 ng 32 P-DNA, 1 μ l UL9 extract, 20 mM HEPES, pH 7.2, 50 mM KCl, and 1 mM DTT). The reactions were incubated at room temperature for 10 minutes (binding occurs in less than 2 minutes), then separated electrophoretically on 4-10% non-denaturing polyacrylamide gels. UL9-specific binding to the oligonucleotide is indicated by a shift in mobility of the oligonucleotide on the gel in the presence of the UL9 protein but not in its absence. Bacterial extracts containing (+) or without (-) UL9 protein and affinity purified UL9 protein were tested in the assay. Only bacterial extracts containing UL9 or affinity purified UL9 protein generate the gel band-shift indicating protein binding.

The degree of extract that needed to be added to the reaction mix, in order to obtain UL9 protein excess relative to the oligonucleotide, was empirically determined for each protein preparation/extract. Aliquots of the preparation were added to the reaction mix and treated as above. The quantity of extract at which the majority of the labelled oligonucleotide appears in the DNA:protein complex was evaluated by band-shift or filter binding assays. The assay is most sensitive under conditions in which the minimum amount of protein is added to bind most of the DNA. Excess protein decreases the sensitivity of the assay with respect to the ability of inhibitors to compete with the protein for oligonucleotide binding, except when protein concentrations are so high that non-specific protein/DNA binding is provoked.

B. Rate of Dissociation.

The rate of dissociation is determined using a competition assay. An oligonucleotide having the sequence presented in Figure 4, which contained the binding site for UL9 (SEQ ID NO:614), was radiolabelled with 32 P-ATP and polynucleotide kinase (Bethesda Research Laboratories). The competitor DNA was a 17 base pair oligonucleotide (SEQ ID NO:616) containing the binding site for UL9.

In the competition assays, the binding reactions (Example 4A) were assembled with each of the oligonucleotides and placed on ice. Unlabelled oligonucleotide (1 μ g) was added 1, 2, 4, 6, or 21 hours before loading the reaction on an 8% polyacrylamide gel (run in TBE buffer (Maniatis, et al.)) to separate the reaction components. The dissociation rates, under these conditions, for the truncated UL9 (UL9-COOH) and the full length UL9 is approximately 4 hours at 4°C. In addition, random oligonucleotides (a 10,000-fold excess) that did not contain the UL9 binding sequence and sheared herring

sperm DNA (a 100,000-fold excess) were tested: neither of these control DNAs competed for binding with the oligonucleotide containing the UL9 binding site.

C. Optimization of the UL9 Binding Assay.

1. Truncated UL9 from the Bacterial Expression System.

The effects of the following components on the binding and dissociation rates of UL9-COOH with its cognate binding site have been tested and optimized: buffering conditions (including the pH, type of buffer, and concentration of buffer); the type and concentration of monovalent cation; the presence of divalent cations and heavy metals; temperature; various polyvalent cations at different concentrations; and different redox reagents at different concentrations. The effect of a given component was evaluated starting with the reaction conditions given above and based on the dissociation reactions described in Example 4B.

The optimized conditions used for the binding of UL9-COOH contained in bacterial extracts (Example 2E) to oligonucleotides containing the HSV ori sequence (SEQ ID NO:601) were as follows: 20 mM HEPES, pH 7.2, 50 mM KCl, 1 mM DTT, 0.005 - 0.1 ng radiolabeled (specific activity, approximately 10^8 cpm/ μ g) or digoxigenated, biotinylated oligonucleotide probe, and 5-10 μ g crude UL9-COOH protein preparation (1 mM EDTA is optional in the reaction mix). Under optimized conditions, UL9-COOH binds very rapidly and has a dissociation rate of about 4 hours at 4°C with non-biotinylated oligonucleotide and 5-10 minutes with biotinylated oligonucleotides. The dissociation rate of UL9-COOH changes markedly under different physical conditions. Typically, the activity of a UL9 protein preparation was assessed using the gel band-shift assay and related to the total protein content of the extract as a method of standardization. The addition of herring sperm DNA depended on the purity of UL9 used in the experiment. Binding assays were incubated at 25°C for 5-30 minutes.

2. Full Length UL9 Protein from the Baculovirus System.

The binding reaction conditions for the full length baculovirus-produced UL9 polypeptide have also been optimized. The optimal conditions for the current assay were determined to be as follows: 20 mM Hepes; 100 mM NaCl; 0.5 mM dithiothreitol; 1 mM EDTA; 5% glycerol; from 0 to 10^4 -fold excess of sheared herring sperm DNA; 0.005 - 0.1 ng radiolabeled (specific activity, approximately 10^8 cpm/ μ g) or digoxigenated, biotinylated oligonucleotide probe, and 5-10 μ g crude UL9 protein preparation. The full length protein also binds well under the optimized conditions established for the truncated UL9-COOH protein.

Example 5

The Effect of Test Sequence Variation on the Half-Life of the UL9 DNA:Protein Complex

The oligonucleotides shown in Figure 5 were radiolabelled as described above. The competition assays were performed as described in Example 4B using UL9-COOH. Radiolabelled oligonucleotides were mixed with the UL9-COOH protein in binding buffer (typical reaction: 0.1 ng oligonucleotide 32 P-DNA, 1 μ l UL9-COOH extract, 20 mM HEPES, pH 7.2, 50 mM KCl, 1 mM EDTA, and 1 mM DTT). The reactions were incubated at room temperature for 10 minutes. A zero time point sample was then taken and loaded onto an 8% polyacrylamide gel (run use TBE). One μ g of the unlabelled 17 bp competitive DNA oligonucleotide (SEQ ID NO:616) (Example 4B) was added at 5, 10, 15, 20, or 60 minutes before loading the reaction sample on the gel. The results of this analysis are shown in Figure 9: the screening sequences that flank the UL9 binding site (SEQ ID NO:605-SEQ ID NO:613) are very dissimilar but have little effect on the off-rate of UL9. Accordingly, these results show that the UL9 DNA binding protein is effective to bind to a screening sequence in duplex DNA with a binding affinity that is substantially independent of test sequences placed adjacent the screening sequence. Filter binding experiments gave the same result.

Example 6

The Effect of Actinomycin D, Distamycin A, and Doxorubicin on UL9 Binding to the screening Sequence

is Dependent on the Specific Test Sequence

Different oligonucleotides, each of which contained the screening sequence (SEQ ID NO:601) flanked on the 5' and 3' sides by a test sequence (SEQ ID NO:605 to SEQ ID NO:613), were evaluated for the effects of distamycin A, actinomycin D, and doxorubicin on UL9-COOH binding.

Binding assays were performed as described in Example 5. The oligonucleotides used in the assays are shown in Figure 5. The assay mixture was allowed to pre-equilibrate for 15 minutes at room temperature prior to the addition of drug.

A concentrated solution of Distamycin A was prepared in dH₂O and was added to the binding reactions at the following concentrations: 0, 1 μ M, 4 μ M, 16 μ M, and 40 μ M. The drug was added and incubated at room temperature for 1 hour. The reaction mixtures were then loaded on an 8% polyacrylamide gel (Example 5) and the components separated electrophoretically. Autoradiographs of these gels are shown in Figure 10A. The test sequences tested were as follows: UL9 polyT, SEQ ID NO:609; UL9 CCCG, SEQ ID NO:605; UL9 GGGC, SEQ ID NO:606; UL9 polyA, SEQ ID NO:608; and UL9 ATAT, SEQ ID NO:607. These results demonstrate that Distamycin A preferentially disrupts binding to UL9 polyT, UL9 polyA and UL9 ATAT.

A concentrated solution of Actinomycin D was prepared in dH₂O and was added to the binding reactions at the following concentrations: 0 μ M and 50 μ M.

The drug was added and incubated at room temperature for 1 hour. Equal volumes of dH₂O were added to the control samples. The reaction mixtures were then loaded on an 8% polyacrylamide gel (Example 5) and the components separated electrophoretically. Autoradiographs of these gels are shown in Figure 10B. In addition to the test sequences tested above with Distamycin A, the following test sequences were also tested with Actinomycin D: AToril, SEQ ID NO:611; oriEco2, SEQ ID NO:612, and oriEco3, SEQ ID NO:613. These results demonstrate that actinomycin D preferentially disrupts the binding of UL9 to the oligonucleotides UL9 CCCG and UL9 GGGC.

A concentrated solution of Doxorubicin was prepared in dH₂O and was added to the binding reactions at the following concentrations: 0 μ M, 15 μ M and 35 μ M. The drug was added and incubated at room temperature for 1 hour. Equal volumes of dH₂O were added to the control samples. The reaction mixtures were then loaded on an 8% polyacrylamide gel (Example 5) and the components separated electrophoretically. Autoradiographs of these gels are shown in Figure 10C. The same test sequences were tested as for Actinomycin D. These results demonstrate that Doxorubicin preferentially disrupts the binding of UL9 to the oligonucleotides UL9polyT, UL9 GGGC, oriEco2, and oriEco3. Doxorubicin appears to particularly disrupt the UL9:screening sequence interaction when the test sequence oriEco3 is used. The sequences of the test sequences for oriEco2 and oriEco3 differ by only one base: an additional T residue inserted at position 12, compare SEQ ID NO:612 and SEQ ID NO:613.

Example 7

Use of the Biotin/Streptavidin Reporter System

A. The Capture of Protein-Free DNA.

Several methods have been employed to sequester unbound DNA from DNA:protein complexes.

1. Magnetic Beads.

Streptavidin-conjugated superparamagnetic polystyrene beads (Dynabeads M-280 Streptavidin, Dynal AS, 6-7x10⁸ beads/ml) are washed in binding buffer then used to capture biotinylated oligonucleotides (Example 1).

The beads are added to a 15 μ l binding reaction mixture containing binding buffer and biotinylated oligonucleotide. The beads/oligonucleotide mixture is incubated for varying lengths of time with the binding mixture to determine the incubation period to maximize capture of protein-free biotinylated oligonucleotides. After capture of the biotinylated oligonucleotide, the beads can be retrieved by placing the reaction tubes in a magnetic rack (96-well plate magnets are available from Dynal). The beads are then washed.

2. Agarose Beads.

Biotinylated agarose beads (immobilized D-biotin, Pierce, Rockford,

IL) are bound to avidin by treating the beads with 50 µg/µl avidin in binding buffer overnight at 4°C. The beads are washed in binding buffer and used to capture biotinylated DNA. The beads are mixed with binding mixtures to capture biotinylated DNA. The beads are removed by centrifugation or by collection on a non-binding filter disc.

For either of the above methods, quantification of the presence of the oligonucleotide depends on the method of labelling the oligonucleotide. If the oligonucleotide is radioactively labelled: (i) the beads and supernatant can be loaded onto polyacrylamide gels to separate DNA:protein complexes from the bead:DNA complexes by electrophoresis, and autoradiography performed; (ii) the beads can be placed in scintillation fluid and counted in a scintillation counter. Alternatively, presence of the oligonucleotide can be determined using a chemiluminescent or colorimetric detection system.

B. Detection of Protein-Free DNA.

The DNA is end-labelled with digoxigenin-11-dUTP (Example 1). The antigenic digoxigenin moiety is recognized by an antibody-enzyme conjugate, anti-digoxigenin-alkaline phosphatase (Boehringer Mannheim Indianapolis IN). The DNA/antibody-enzyme conjugate is then exposed to the substrate of choice. The presence of dig-dUTP does not alter the ability of protein to bind the DNA or the ability of streptavidin to bind biotin.

1. Chemiluminescent Detection.

Digoxigenin-labelled oligonucleotides are detected using the chemiluminescent detection system "SOUTHERN LIGHTS" developed by Tropix, Inc. (Bedford, MA). Use of this detection system is illustrated in Figures 11A and 11B. The technique can be applied to detect DNA that has been captured on either beads or filters.

Biotinylated oligonucleotides, which have terminal digoxigenin-containing residues (Example 1), are captured on magnetic (Figure 11A) or agarose beads (Figure 11B) as described above. The beads are isolated and treated to block non-specific binding by incubation with I-Light blocking buffer (Tropix) for 30 minutes at room temperature. The presence of oligonucleotides is detected using alkaline phosphatase-conjugated antibodies to digoxigenin. Anti-digoxigenin-alkaline phosphatase (anti-dig-AP, 1:5000 dilution of 0.75 units/ul, Boehringer Mannheim) is incubated with the sample for 30 minutes, decanted, and the sample washed with 100 mM Tris-HCl, pH 7.5, 150 mM NaCl. The sample is pre-equilibrated with 2 washes of 50 mM sodium bicarbonate, pH 9.5, 1 M MgCl₂, then incubated in the same buffer containing 0.25 mM 3-(2'-spiroadamantane)-4-methoxy-4-(3'-phosphoryloxy) phenyl-1,2-dioxetane disodium salt (AMPPD) for 5 minutes at room temperature. AMPPD was developed (Tropix Inc.) as a chemiluminescent substrate for alkaline phosphatase. Upon dephosphorylation of AMPPD the resulting compound decomposes, releasing a prolonged, steady emission of light at 477 nm.

Excess liquid is removed from filters and the emission of light occurring as a result of the dephosphorylation of AMPPD by alkaline phosphatase can be measured by exposure to x-ray film or by detection in a luminometer.

In solution, the bead-DNA-anti-dig-AP is resuspended in "SOUTHERN LIGHT" assay buffer and AMPPD and measured directly in a luminometer. Large scale screening assays are performed using a 96-well plate-reading luminometer (Dynatech Laboratories, Chantilly, VA). Subpicogram quantities of DNA (10² to 10³ attomoles (an attomole is 10⁻¹⁸ moles)) can be detected using the Tropix system in conjunction with the plate-reading luminometer.

2. Colorimetric Detection.

Standard alkaline phosphatase colorimetric substrates are also suitable for the above detection reactions. Typically substrates include 4-nitrophenyl phosphate (Boehringer Mannheim). Results of colorimetric assays can be evaluated in multiwell plates (as above) using a plate-reading spectrophotometer (Molecular Devices, Menlo Park CA). The use of the light emission system is more sensitive than the colorimetric systems.

Example 8
Labelling Test Oligonucleotides to
Equivalent Specific Activities

The top strands of 256 oligonucleotides, containing all possible 4 bp sequences in the test sites flanking the UL9 recognition site, were synthesized. The oligonucleotides were composed of identical sequences except for the 4 bp sites flanking either side of the UL9 recognition sequence (SEQ ID No:601). The oligonucleotides had the general sequence presented in Figure 14B (SEQ ID NO:617), where XXXX is the test sequence and N = A,G,C, or T. A 12 bp primer sequence, which is the complementary sequence to the 3'-end of the test oligonucleotide, was also synthesized: the primer was designated the HSV primer and is presented as SEQ ID NO:618.

The HSV primer was used to prime second strand synthesis and to facilitate labeling the oligonucleotides to the same specific activity. Oligonucleotide labelling was accomplished by labeling the 5' end of the HSV primer and then using the same primer to prime second strand synthesis of all 256 test oligonucleotides. The 5' end of the primer can be labeled with radioisotopes such as ^{32}P , ^{33}P , or ^{35}S , or with non-radioactive detection systems such as digoxigenin or biotin as discussed in the Capture/Detection section.

Radioactive-labeling of the primer with ^{32}P is accomplished by the enzymatic transfer of a radioactive phosphate from $\gamma\text{-}^{32}\text{P}\text{-ATP}$ to the 5' end of the primer oligonucleotide using T4 polynucleotide kinase (Ausubel, et al.). For labeling 256 oligonucleotides, approximately 60 μg HSV primer was labeled as follows. The oligonucleotide was incubated for 1 hour at 37°C with 125 μl $\gamma\text{-}^{32}\text{P}\text{-ATP}$ (20 mCi total, 7000 Ci/mmol) and 600 units of T4 polynucleotide kinase in a 3 ml reaction volume containing 50 mM Tris-HCL, pH 7.5, 10 mM MgCl_2 , 10 mM spermidine, and 1.5 mM dithiothreitol (freshly prepared). To stop the reaction, EDTA was added to a final concentration of 20 mM. Unincorporated nucleotides were removed using "G-25 SEPHADEX" chromatography in 10 mM Tris-HCL, pH 7.5, 50 mM NaCl, and 1 mM EDTA (TE+50).

The radioactive primer was individually annealed to the top strand of each of the 256 test oligonucleotides. The bottom strand is synthesized using deoxyribonucleotides and Klenow fragment or T4 polymerase (Ausubel, et al.). The annealing mixture typically contained 200 ng HSV primer mixed with 1 μg top strand in 20 mM Tris-HCL, pH 7.5, 1 mM spermidine, and 0.1 mM EDTA (35 μl reaction volume). The primer was annealed to the top strand by incubating the sample for 2 minutes at 70°C , then placing the sample at room temperature or on ice. To the annealing mixture, 4.5 μl 10x Klenow buffer (10X = 200 mM Tris-HCL, 500 mM NaCl, 50 mM MgCl_2 , 10 mM dithiothreitol), 5 μl 0.5 mM each dNTP (dATP, dCTP, dGTP, dTTP), and 1 μl Klenow fragment were added. This reaction mixture was incubated 30-60 minutes at room temperature (or up to 37°C).

The volume of the reaction mixture was increased by adding 75 μl a solution of 10 mM Tris-HCL, pH 7.5, 50 mM NaCl, and 10 mM EDTA. The reaction mixture was applied to a 1 ml "G-25 SEPHADEX" (in TE+50) spin column. The spin columns were prepared by plugging lcc tuberculin syringes with silanized glass wool and adding a slurry of "G-25 SEPHADEX." The columns were prespun at 2000 rpm in a tabletop centrifuge for 4 minutes. The samples (reaction mixtures) were passed through the column by centrifugation (2000 rpm, 4 minutes at room temperature) to remove unincorporated deoxyribonucleotides. The incorporation of ^{32}P was measured by placing a small volume of the sample in scintillation fluor and determining the disintegrations per minute (dpms) in a scintillation counter.

The radiolabeled double-stranded oligonucleotides were then diluted to the same specific activity (equal dpms per volume). Typically, a concentration of 0.1 to 1 ng/ μl oligonucleotide was used in the assay.

The same procedure can be used for second strand synthesis and labeling to equal specific activity regardless of the type of label on the HSV primer.

Example 9**An Arrayed Sample Format**

Screening large numbers of test molecules or test sequences is most easily accomplished in an arrayed sample format, for example, a 96-well plate format. Such formats are readily amenable to automation using robotics systems. Several different types of disposable plastic plates are available for use in screening assays including the following: polyvinyl chloride (PVC), polypropylene (PP), polyethylene (PE), and polystyrene (PS) plates. Plates, or any testing vehicle in which the assay is performed, are tested for protein and DNA adsorption and coated with a blocking reagent if necessary.

One method for testing protein or DNA adsorption to plates is to place assay mixtures in the wells of the plates for varying lengths of time. Samples are then removed from the wells and a nitrocellulose dot blot capture system (Ausubel, et al.; Schleicher and Schuell) is used to measure the amount of DNA:protein complex remaining in the mixture over time.

When radiolabeled oligonucleotides are used for the test, signal can be measured using autoradiography and a scanning laser densitometer. A decrease in the amount of DNA:protein complex in the absence of competitor molecules is indicative of plate adsorption. If plate adsorption occurs, the plates are coated with a blocking agent prior to use in the assay.

None of the plates listed above showed marked adsorption at a 30 minute time point under the conditions of the assay. However, most plates, regardless of brand, showed significant adsorption at times greater than 2 hours.

Coating the plates with a blocking agent decreases variability in the assay. Several types of blocking reagents typically used to block the adsorption of macromolecules to plastic are known, primarily from immunoscreening procedures. For example, plates may be blocked with either 1% bovine serum albumin (BSA) in phosphate-buffered saline (PBS), or 0.1% gelatin, 0.05% "TWEEN29" in PBS.

To test for the effectiveness of using such blocking reagents, the plates were treated with the above reagents for 1 hour at room temperature, then washed three times with 0.05% "TWEEN20" in PBS and once with the assay buffer.

Assay reaction mixtures were aliquoted to the plates and tested as described above using dot blot capture assays. Both of the blocking reagents (BSA or gelatin) were effective in blocking DNA and protein binding -- except when polypropylene plates were used. Based on these experiments, PVC plates blocked with BSA were determined to work well in the assay of the present invention.

Plates were tested for inter- and intra-plate variability by aliquoting duplicate samples to all 96-wells of several plates, and determining the amount of DNA:protein complex recovered using the dot blot/nitrocellulose system. The coefficient of variation [%CV = (the standard deviation/mean)*100] was calculated for intra-plate variability (i.e., between samples on the same plate) and inter-plate variability (i.e., between plates). Blocked PVC plates showed an intra-plate %CV of 5-20%; inter-plate variability was about 8%.

Example 10**Sequence Selectivity and Relative Binding Affinity for Distamycin**

using the assay method of the present invention, distamycin was tested for sequence selectivity and relative binding affinity to 256 different 4 bp sequences.

A. The Assay Mixture.

Water, buffer and UL9 were mixed on ice and aliquoted to the wells of a 96-well plate. The addition of water/UL9/buffer mix was accomplished with an 8-channel repipettor, which holds a relatively large volume and allowed rapid, accurate pipetting to all 96 wells of a master experimental plate.

Radiolabeled double-stranded oligonucleotides were aliquoted from 96-well master stock plates (containing the array of all 256 oligonucleotides diluted to the same specific activity) to the wells of the master experimental plates.

Master assay mixtures in the master experimental plates were thoroughly mixed by pipetting up and down. The mixtures were aliquoted to the test plates. Each test plate typically included one sample as a control (no test molecules added) and as many test samples as were needed for different test molecules or test molecule concentrations. There were 3 master oligonucleotide

stock plates, containing the array of 256 oligonucleotides. Accordingly, an experiment testing distamycin at different concentrations would require 256 control assays (one for each oligonucleotide) and 256 assays at each of the drug concentrations to be tested.

The following assay mixture was used for testing distamycin in the assay of the present invention: 1.5 nM radiolabeled DNA and 12.8 nM UL9-COOH protein (prepared as described above in the UL9 binding buffer; 20 mM Hepes, pH 7.2, 50 mM KCl, and 1 mM dithiothreitol). The concentration of the components in the assay mixture can be varied as described above in the Detailed Description.

Assay mixtures containing both UL9 and DNA were incubated at room temperature for at least 10 minutes to allow the DNA:protein complexes to form and for the system to come to equilibrium. At time = 0, the assay was begun by adding water (control samples) or distamycin (5-15 μ M, test samples) to the assay mixtures using a 12-channel micropipettor. After incubation with drug for 5-120 minutes, samples were taken and applied to nitrocellulose on a 96-well dot blot apparatus (Schleicher and Schuell). The samples were held at 4°C.

Tests were performed in duplicate. Typically, one set of 256 test oligonucleotides was scrambled with respect to location on the 96-well plate to eliminate any effects of plate location.

B. The Capture/Detection System.

A 96-well dot blot apparatus was used to capture the DNA:protein complexes on a nitrocellulose filter. The filters used in the dot blot apparatus were pretreated as follows. The nitrocellulose filter was pre-wetted with water and soaked in UL9 binding buffer. The filter was then placed on 1 to 3 pieces of 3MM filter paper, which were also presoaked in UL9 binding buffer. All filters were chilled to 4°C prior to placement in the apparatus.

Prior to the application of the assay sample to the wells of the dot-blot apparatus, the wells were filled with 375 μ l of UL9 binding buffer. Typically, 5-50 μ l of sample (usually 10-15 μ l) were pipetted into the wells containing binding buffer and a vacuum applied to the system to pull the sample through the nitrocellulose. Unbound DNA passes through the nitrocellulose, protein-bound DNA sticks to the nitrocellulose. The filters were dried and exposed to X-ray film to generate autoradiographs.

C. Quantitation of Data.

The autoradiographs of the nitrocellulose filters were analyzed with a Molecular Dynamics (Sunnyvale, CA) scanning laser densitometer using an ImageQuant software package (Molecular Dynamics). Using this software, a 96-well grid was placed on the image of the autoradiograph and the densitometer calculated the "volume" of each dot ("volume" is equivalent to the density of each pixel in the grid square multiplied by the area of the grid square). The program automatically subtracts background. The background was determined by either the background of a line or object drawn outside the grid or by using the gridlines as background for each individual dot.

The data is exported to a spreadsheet program, such as "EXCEL" (Microsoft Corporation, Redmond, WA) for further analysis.

D. Analysis of Data.

The data generated from the densitometry analysis was analyzed using the spreadsheet program "EXCEL."

For each test oligonucleotide, at each drug concentration and/or each time point, a raw % score was calculated. The raw % score (r%) can be described as

$$r\% = (T/C) \times 100$$

where T was the densitometry volume of the test sample and C was the densitometry volume of the control sample. The oligonucleotides were then ranked from 1 to 256 based on their r% score. Further calculations were based on the rank of each oligonucleotide with respect to all other oligonucleotides.

The rank of each oligonucleotide was averaged over several experiments

(where one experiment is equivalent to testing all 256 test oligonucleotides by the assay of the present invention) in view of the variability in rank between any two experiments. The confidence level for the ranking of the oligonucleotides increased with repetition of the experiment.

Figure 15 shows the results of 4 separate experiments with distamycin. The test samples were treated with 10 μ M distamycin for 30 minutes. The $r\%$ scores are shown for each of the 4 experiments (labeled 918A, 918B, 1022A, and 1022B) and the ranks of each oligonucleotide in each experiment are shown. The test oligonucleotides have been ranked from 1 to 256 based on their average rank. The average rank was the sum of the ranks in the individual experiments divided by the number of experiments.

Figures 16 and 17 show the results presented in Figure 15 in graphic form. Figure 16 shows the average ranks plotted against the ideal ranks 1 to 256. Figure 17 shows the average $r\%$ scores plotted against the rank of 1 to 256. These data demonstrate the reproducible ability of the assay to detect differential binding and effects of distamycin on different 4 bp sequences.

Example 11

Determining a Consensus Binding Site for Distamycin

One method used to determine the sequence preferences for distamycin was to examine the sequences that rank highest in the assay for sequence similarities. This process may be accomplished visually or by designing computer programs to inspect the data.

Using the data shown in Figure 15, consensus sequences can be constructed for distamycin in the following manner. Sequences with rankings less than 50 (indicating a strong effect of distamycin on the test sequence) in all four experiments were:

TABLE VI

Sequence	Rank
TTCC	1
TTAC	2
TACC	3
TATC	4
TTCG	6
ACGG	8

Sequences with rankings less than 50 (indicating a strong effect of distamycin on the test sequence) in three of the four experiments were:

TABLE VII

Sequence	Rank
AACG	5
TTTC	7
TTAG	10
TAAC	12
TACG	15
AGAC	17
AAAC	18
AGCG	21
AGCC	22
TTCT	24
ACGC	25
AGGG	28
AGGC	30
TTGC	37
ATCG	39
TTTG	43

Sequences with rankings less than 50 (indicating a strong effect of distamycin on the test sequence) in two of the four experiments were:

TABLE VIII

Sequence	Rank
TAGC	9
TTGG	11
AAAG	13
AACC	14
CAAC	16
ATCC	19
AAGG	20
TAAG	23
ACCC	26
TCCC	29
TATG	31
ACCG	32
TCGG	34
AGTC	35
CTCG	38
AATC	44
AGAG	46
TTAA	47
ACAC	48
AGTG	49
TCAC	52

The following assumptions allow prediction of a consensus sequence for a distamycin recognition sequence: (i) the most favored sequences are the test sequences that rank in the top 50 in all four experiments; (ii) the next favored sequences will be the test sequences that rank in the top 50 in 3 of 4 experiments; and (iii) the next favored sequences will be the test sequences that rank in the top 50 in 2 of 4 experiments.

The positions in the test sequence are represented by the numerals 1, 2, 3 and 4. One consensus sequence that predicted from the above binding data is:

1	2	3	4
T	T/A	N	C/G

The nucleotides at each position can also be ranked:

1	2	3	4
T	T>A	C>A>T>G	C>G

Furthermore, the importance of the position of the nucleotide can be ranked. Examination of this data would indicate that the importance of the positions is

1 > 4 > 2 > 3.

These data can be tested for validity by deriving all possible consensus sequences and examining their scores in the assay. The consensus sequences derived from the above information, in order of rank as predicted by the consensus sequence, are:

TABLE IX

Sequence	Predicted Rank	Actual Rank
TTCC	1	1
TACC	2	3
TTCG	3	6
TACG	4	15
TTAC	5	2
TAAC	6	12
TTAG	7	10
TAAG	8	23

Sequence	Predicted Rank	Actual Rank
TTTC	9	7
TATC	10	4
TTTG	11	43
TATG	12	31
TTGC	13	37
TAGC	14	9
TTGG	15	11
TAGG	16	58
Average rank:		17

Note that the actual rank numbers are out of a possible 256 and that only one number is greater than 50. The average rank of these 16 oligos is only 17. These data indicate that the consensus sequence has predictive value.

- 5 Using the same data, a second consensus sequence can be derived that has slightly worse average rank with respect to the relative effect of distamycin in the assay.

TABLE X

1	2	3	4
A	A/G/C	G/C/A	G/C
	A>G=C	C>A=G	G=C

10

The test sequences predicted by this consensus sequence are as follows:

TABLE XI

Sequence	Actual rank
AACG	5
AACC	14
AAAG	13
AAAC	18
AAGG	20
AAGC	74
AGCG	21
AGCC	22
AGAG	46
AGAC	17
AGGG	28
AGGC	30
ACCG	32
ACCC	26
ACAG	73
ACAC	48
ACGG	8
ACGC	25
Ave. rank:	29

- 15 This consensus sequence also appears to be predictive of favored distamycin binding sites since the average rank of test oligonucleotides predicted by this sequence is 29, substantially below the median rank of 128. However, the sequences predicted by this consensus sequence do not appear to be affected as strongly by distamycin as the sequences in the first consensus sequence, described above.
- 20

Example 12Testing Actinomycin D to Determine SequenceSpecificity and Relative Binding AffinityA. Ranking of Actinomycin D Sequence Binding Affinities.

Actinomycin D has been tested for sequence selectivity and relative binding affinity to the 256 different 4 bp sequences. The assay was performed essentially as described in Example 10. One assay mixture useful for the testing of actinomycin D contained 1.5 nM radiolabeled DNA and 12.8 nM UL9-COOH protein prepared as described above in the UL9 binding buffer (20 mM Hepes, pH 7.2, 50 mM KCl, and 1 mM dithiothreitol). The concentration of the components can be varied as described in the Detailed Description. The assay mixtures containing both UL9 and DNA were incubated at room temperature for at least 10 minutes to allow the DNA:protein complexes to form and for the system to come to equilibrium. At time = 0, the assay was begun by adding water (control samples) or actinomycin D (25 μ M, test samples) to the assay mixtures using a 12-channel micropipettor. After incubation with drug for 30 minutes, samples were taken and applied to nitrocellulose filters using a 96-well dot blot apparatus (Schleicher and Schuell) held at 4°C. Figure 18 shows the results of 8 screens of actinomycin D.

The % reduction in DNA:protein complex as a result of the presence of actinomycin D is called "r%"; the lower the r% score, the more effective the test molecule in blocking the DNA:protein interaction. For each screen, the test oligonucleotides have been ranked from 1 to 256, based on the r% score; the rank of 1 denotes the lowest r% score (the test oligonucleotide most effected by the test molecule), the rank of 256 denotes the highest r% score (the test oligonucleotide least effected by the test molecule). The table also shows the average r% score and average rank of each test oligonucleotide; the averages are calculated from the sum of the individual scores and ranks divided by the number of screens, respectively. The test oligonucleotides are then ranked from 1 to 256 based on the average rank in all screens. The final ranking is shown in the two external columns on the table. Test oligonucleotides ranking less than 50 in any individual screen are shown in highlighted boxes.

Figure 19 shows the final rank of test oligonucleotides screened with actinomycin D plotted against the average r% score for these test oligonucleotides.

Figure 20 shows the final ranking vs. the ranks in each individual experiment, the average rank, and the ideal rank.

B. Analysis of the Data Obtained from Ranking Actinomycin D Sequence Binding Affinities.

Several simple analytical procedures may be applied to the data from the screens.

1. Position Effects.

First, to examine possible preferences of the test molecule for a base at any particular position in the test site, the average r% scores are examined. The average r% scores for each of the 64 possible test oligonucleotides at each position in the test site are averaged. For example, to determine the effect of having an A in the first position of the test site, the "A₁" position, the average r% scores for the 64 test oligonucleotides with A in the first position are averaged. The results of this analysis are shown in Figure 21. The mean score for all oligonucleotides in these screens was r% value 67; the standard deviation was 11.8.

If the r% score is expressed as variance from the mean, as shown in Figure 21, one observes that none of the scores is markedly deviant from the mean. These results suggest that a single base in any particular position has little impact on the binding of the actinomycin D to the test site.

2. Dinucleotide Analysis.

The results of the actinomycin D screen were examined for the presence of dinucleotide pairs that scored well or poorly in the rankings. High scores indicate a preference for the test sequence. Low scores indicate a repulsion of actinomycin D for the test sequence. A dinucleotide analysis is

one of many simple analytical procedures that may be applied to the data to extract meaningful impressions about the nature of the sequences to which the test molecule has high affinity.

The data are examined in a manner similar to that used for the single nucleotide analysis. The 16 possible average $r\%$ scores for any particular dinucleotide combination are examined. Specific adjacent dinucleotides (N_1N_2 , N_2N_3 , N_3N_4) or adjacent dinucleotide pairs at any particular position (N_xN_{x+1} = the average of N_1N_2 , N_2N_3 , and N_3N_4) may be examined, as well as specific dinucleotide pairs that are not adjacent (N_1N_3 , N_2N_4 , N_1N_4) and any dinucleotide pair separated by one base (N_xN_{x+2} = the average of N_1N_3 and N_2N_4). The means for each set are determined as well as standard deviations.

The difference from the mean (i.e., the mean score less the average $r\%$ score for any particular dinucleotide) reflects the extent of deviation from the norm. Differences from the mean greater than 2-3 standard deviations from the mean are considered to be significant. The data for the dinucleotide analysis of actinomycin D is shown in Figure 22. The differences from the mean are displayed graphically in Figure 23.

In reference to Figures 22 and 23, the dinucleotide preference of actinomycin D is GC, particularly in the N_1N_2 position, but also at any (N_xN_{x+1}) adjacent dinucleotide sequence in the test site.

If the data are combined in a combined bar chart, shown in Figure 24, where the cumulative results for any dinucleotide pair are tabulated in a single bar, the overall observation can be made that actinomycin D prefers GC-rich sequences over AT-rich sequences, with a particular preference for the dinucleotide pairs involving GC.

Example 13

A Method for Selecting Target Sites for DNA-Binding Molecules that are Dimers or Trimers of Distamycin

Once the relative binding preferences of a distamycin have been determined, sequences are selected for target sites for DNA-binding molecules composed of two distamycin molecules, bis-distamycins, or three distamycin molecules, tris-distamycins.

A. Selecting Sequences for Binding with Highest Affinity to Distamycin Oligomers.

The top binding sites for distamycin, determined as described above, are defined by the consensus sequence, 5'-T:T/A:C/A:C-3': accordingly, the top sequences are TTCC, TTAC, TACC and TAAC. Using this information, $2^4 = 16$ possible dimer sequences, i.e., combinations of the four top binding sequences, can be targeted by a bis-distamycin in which the distamycin molecules are immediately adjacent to one another.

The top strands of the 16 possible duplex DNA target sites for binding bis-distamycins are shown in Figure 25. Similarly, trimers of distamycin, tris-distamycins, could be targeted toward selected 12 bp sequences, comprised of all possible combinations of the four 4 bp sequences. There are $3^4 = 81$ possible highest affinity target trimer sequences. There are several advantages to targeting longer sequences with bis- or tris-distamycin:

B. As the Number of Potential Target Sites Decreases, Specificity Increases.

All 8 bp combinatorial possibilities of the 4 top favored binding sites for distamycin are potential high affinity binding sites for bis-distamycin. The consensus sequence used in this example predicts four favored binding sites for distamycin. This represents $(4/4^4) \times 100 =$ about 1.6% of the possible 4 bp sites in the genome. Since there are 4^8 possible 8 bp sequences, this represents, on average, only $(2^4/4^8) \times 100 =$ about 0.02% of the total genome. There are 4^{12} possible 12 bp sequences, this represents, on average, only $(3^4/4^{12}) \times 100 = 0.00000075\%$ of the genome.

The following discussion provides perspective and illustrates the improvement in the actual number of target sites in the human genome for when using a dimer of distamycin versus a monomer of distamycin. The human genome is about 3×10^9 bp. If the number of favored target sites for distamycin is

four, and the number of possible 4 bp sequences is $4^4 = 256$, then the number of favored target sites in the genome is $(4/256)(3 \times 10^9) = 4.7 \times 10^7$, or about 50 million favored target sites.

Given that the number of possible 8 bp sites is $4^8 = 65,536$, if all possible combinatorial 8 bp sites derived from the favored 4 bp sites ($2^4 = 16$; Figure 25) are favored, then the number of favored 8 bp target sites is $(16/65,536)(3 \times 10^9) = 7.3 \times 10^5$ or about 700,000 possible sites. This represents a 64-fold reduction in the number of highest affinity target sites between distamycin and bis-distamycin; alternatively, this result can be viewed as a 64-fold increase in specificity.

Likewise, given that the number of possible 12 bp sites is $4^{12} = 1.7 \times 10^7$, if all possible favored 12 bp sites ($3^4 = 81$) are favored, then the number of favored 12 bp target sites is $(81/1.7 \times 10^7)(3 \times 10^9) = 1.4 \times 10^4$: i.e., 14,000 possible highest affinity sites. This represents an approximately 3000-fold decrease in the number of highest affinity target sites between distamycin and tris-distamycin and a 500-fold decrease in the number of highest affinity target sites between bis-distamycin and tris-distamycin.

C. An Exponential Increase in Affinity.

As the target site increases in size, (i) the number of target sites in a defined number of nucleotides decreases, and (ii) the specificity increases. Further, the affinity of binding is typically the product of the binding affinities of component parts (see Section VI.E.1 above). As an example, the published binding constant for distamycin to bulk genomic DNA is about $2 \times 10^5 \text{ M}^{-1}$. Dimers of distamycin will have a theoretical binding affinity of the square of the binding constant of distamycin:

$$(K_{\text{dista, average}} = 2 \times 10^5 \text{ M}^{-1}; K_{\text{bis-dista}} = (2 \times 10^5 \text{ M}^{-1})^2 = 4 \times 10^{10} \text{ M}^{-1}).$$

Trimers of distamycin will have binding affinities of the cube of the binding affinity of distamycin:

$$(K_{\text{tris-dista}} = (2 \times 10^5 \text{ M}^{-1})^3 = 8 \times 10^{15} \text{ M}^{-1}).$$

Thus, if distamycin shows only a 10-fold higher affinity ($2 \times 10^6 \text{ M}^{-1}$) for the top favored binding sites than the average binding sites in DNA, then the affinity constant for bis-distamycin to an 8 bp site comprised of two favored binding sites is 100-fold higher than for an 8 bp sequence comprised of two average binding sites:

$$(K_{\text{bis-dista, favored sites}}/K_{\text{bis-dista, average sites}} = (2 \times 10^6)^2/(2 \times 10^5)^2 = 100). \text{ While this does not represent absolute sequence specificity in binding, the binding affinity is 100-fold greater for 0.02\% (16/65,536) of the total possible 8 bp target sequences.}$$

The use of a trimer targeted sequence will afford an even higher increase in affinity to the most favored binding sites:

$$K_{\text{tris-dista, favored sites}}/K_{\text{tris-dista, average sites}} = (2 \times 10^6)^3/(2 \times 10^5)^3 = 1000. \text{ Thus, with only 10-fold differential activity in binding between favored sites and average sites, a 1000-fold difference in affinity can be achieved by designing trimer molecules to specific target sites. When considering the administration of DNA-binding molecules as drugs, a 1000-fold lower dose of tris-distamycin, versus the distamycin monomer, could be administered and an increase in relatively specific binding to selected target sites achieved.}$$

In this example, the differential activity of distamycin is only 10-fold. Clearly, differential activities of larger magnitudes will greatly accentuate the increased affinity effect. For example, a 100-fold difference in activity of a 4 bp DNA-binding molecule toward high affinity and average affinity sequences would result in (i) a 10,000-fold difference in the binding affinity of a dimer of the molecule targeted to an 8 bp sequence, and (ii) a million-fold increase in the binding affinity of the trimer to a 12 bp sequence.

D. Selecting Target Sequences for Distamycin Oligomers with Flexible and/or Variable-Length Linkers in Between the Distamycin Moieties.

The sequences that can be targeted with bis- or tris-distamycin molecules are not limited to sequences in which the two 4 bp favored binding sites are immediately adjacent to one another. Flexible linkers can be placed between the distamycin moieties and sequences can be targeted that are not immediately adjacent. The target sequences can have distances of 1 to several bases between them: this distance depends on the length of the chemical linker. Examples of bis-distamycin target sequences for bis-distamycins with internal flexible and/or variable length linkers targeted to sites comprised of two TTCC sequences are shown in Figure 26, where N is any base.

For each particular bis-distamycin, the explanations of increased affinity and specificity remain the same as described above with the following exception. For the case in which the linker was sufficiently flexible to span different numbers of bases in between the two distamycin sites, the number of sites targeted with highest affinity would be multiplied by the number of bases spanned.

In respect to the ease of drug design and target selection, there are several advantages to the above described targeting strategies, including the following:

i) Any conformational changes induced by binding at the half-site would be minimized.

ii) The affinity, therefore, would be more likely to be the product of the affinities of the interactions observed for the monomeric sites.

iii) The half-molecule (e.g., 1 distamycin unit) would anchor the bis-molecule (e.g., bis-distamycin) thus increasing the localized concentration for the binding of the second half of the bis-molecule.

iv) If a simple linking chain is used, with a variable number of atoms, the number of sites that can be targeted by multimers of the monomer increases.

This targeting method can be of value when, for example, there are no medically significant target sites with adjacent favored binding sites for distamycin. Therefore there are no good target sites for bis-distamycin. In this situation, the database can be screened for additional target sequences with N_1 to n (where N is any base) between the two target binding sequences. For example, where $n=4$, the number of sequences to be searched becomes $(4^2) \times 4 = 64$. The likelihood of finding such a sequence is reasonably high.

E. Selecting a Specific Target Site.

Using the above approach, a sequence was identified from the medically significant target site database that contains SEQ ID NO:619, which is a subset of the group of sequences represented by SEQ ID NO:620. SEQ ID NO:619 occurs overlapping the binding site for a transcription factor, Nuclear Factor of Activated T Cells (NFAT-1), which is a major regulatory factor in the induction of interleukin 2 expression early in the T cell activation response. NFAT-1 is crucial in (i) the T cell response, and (ii) in blocking the expression of IL-2, which causes immunosuppression. The sequences TTCC and TTTC, the distamycin target binding sequences in SEQ ID NO:619, rank first and seventh in the assay.

Example 14

The Use of the Assay in Competition Studies

The assay of the present invention measures the effect of the binding of a DNA-binding molecule to a test site by the release of a protein from an adjacent screening site. Accordingly, the assay is an indirect assay. Following here is the description of an application of the assay useful to provide confirmatory evidence of the data obtained in the initial screening processes.

The results of the distamycin screening assay described in Example 10 suggested that there were possible false negatives: specifically, test sequences that bind distamycin but fail to show an effect on the binding of the reporter protein. The data suggesting false negatives was as follows. If the assay detected strictly the affinity of binding of distamycin, then the scores of the test sequences complementary to the high-scoring test sequences should always be equally high. However, an examination of the highest ranking test sequences and the complementary test sequences reveals that this is not the

case (see Table XII).

TABLE XII

Rank	Test Sequence	Complement	Rank of complement
1	TTCC	GGAA	42
2	TTAC	GTAA	244
3	TACC	GGTA	185
4	TATC	GATA	213
5	AACG	CGTT	144
6	TTCG	CGAA	216
7	TTTC	GAAA	235

All but one of the complementary sequences rank in the lower half, 4 of them in the lowest 20%, i.e., these had little effect on reporter protein binding in the presence of distamycin when using these sequences as test sequences in the assay.

This observation reflects the usefulness of a confirmatory assay that examines the relative affinity of a particular sequence for binding distamycin.

A confirmatory assay may also be useful in revealing additional information about the physical characteristics of drug binding. For example, one can hypothesize that the reason for the apparent inverse relationship between test sequences with high activity in the assay and their complements is that the effect of distamycin is directional and only active at one test site. This hypothesis can be tested using the following competition experiment. Competitor oligonucleotides, containing test sequences of interest, are added to the assay mixture. This allows the determination of which test sequences compete most effectively with the radiolabeled test oligonucleotide for binding distamycin.

Assay mixtures are prepared as described in Example 10, using a high-ranking test oligonucleotide, e.g., TTCC (ranking = #1), as the radiolabeled oligonucleotide in the experiment. The test oligonucleotide TTCC is labeled to high specific activity with γ -³²P-ATP as described in Example 8: in this example, the labeled TTCC oligonucleotide will be referred to as the "high specific activity test oligonucleotide".

The competitor oligonucleotides are labeled as described in Example 8, except that the ATP used for kinasing the primer is 1:200 radiolabeled:nonradiolabeled. In other words, the competitor oligonucleotides are tracer labeled with radioactive phosphorous to a 200-fold lower specific activity than the high specific activity test oligonucleotide. Since all of the competitor oligonucleotides are labeled with the same radiolabeled primer molecule, the relative concentrations of the competitor DNAs can be determined with high accuracy. Further, since the specific activity is the same, the concentrations can be adjusted to be the same. For the purposes of this example, the competitor DNAs are referred to as "low specific activity competitor oligonucleotides."

The use of competitor DNAs for which the concentration is known is important for the competition experiment. The accuracy of the competition assay may be further enhanced by separating any unincorporated radiolabeled primer from the double stranded competitor oligonucleotides. This separation can be achieved using, for example, a 6-20% polyacrylamide gel. The gel is then exposed to x-ray film and the amount of double-stranded oligonucleotide determined by use of a scanning laser densitometer, essentially as described in the Examples above.

The competition assay is performed as described in Example 10, except that competitor DNAs are added in increasing relative concentration to the high specific activity test oligonucleotide. The DNA concentration ([DNA]) is held constant and the UL9 concentration ([UL9]) and distamycin concentration ([distamycin]) are as described in Example 10. The components in the competition assay samples are as follows.

Controls:

UL9 + TTCC*; UL9 + TTCC* + Competitors; UL9 + TTCC* + distamycin;

Test samples:

UL9 + TTCC* + distamycin + Competitors;

where UL9 is UL9-COOH, TTCC* is the high specific activity test oligonucleotide, and Competitors are the low specific activity competitor oligonucleotides.

TTCC-low (the tracer-labeled low specific activity competitor) competes with TTCC* on an equimolar basis for the binding of both protein and distamycin. A competitor molecule with lower affinity for distamycin than TTCC requires a higher molar ratio to TTCC* to compete for distamycin binding. The competition for protein between all competitors is constant. Only the competition for distamycin varies; the variability is due to the differential affinity of the competitor oligonucleotides for distamycin. The concentration of competitor used in these experiments varies over a range of concentrations and is determined empirically by (a) the test molecule concentration, and (b) the relative affinity of the competitor and the radiolabeled test oligonucleotide. Typically, the competitor DNA consists of only the test sequence, that is, no additional sequences are connected to the test sequence.

The competition assay described here facilitates the determination of actual rank between the test oligonucleotides that are detected as highly effective molecules in the original assay. The competition assay also facilitates the detection of false negatives. As described above, the results of the assay discussed in Example 10 imply "directional" binding of distamycin, in which the effect of binding is only detected when the molecule is bound in one direction with respect to the UL9 protein. Binding in the opposite direction (i.e., to the complementary test sequence) is not detected with the same activity in the assay.

The purpose of this competition experiment is to use the test oligonucleotides to compete for the binding of distamycin. If the sequences complementary to the "best binders" are false negatives in the assay, they should nonetheless be effective competitors in the competition assay.

Example 15

A Method of Selecting Target Sequences From Database Sequence Information

The binding of a drug or other DNA-binding molecule to the recognition sequence for TFIID, or other selected transcription factors, is expected to alter the transcriptional activity of the associated gene. TATA-boxes, which are the recognition sequences for the transcriptional regulatory factor TFIID, are associated with most eukaryotic promoters and are critical for the expression of most eukaryotic genes. Targeting a DNA-binding drug to TATA boxes in general would be undesirable. However, sequences flanking TATA box sequences are typically unique between genes. By targeting such flanking sequences, perhaps with one base overlapping the TFIID recognition site, each gene can be targeted with specificity using the novel DNA-binding molecules designed from the data generated from the DNA-binding drug assay. One method for determining novel and specific target sequences for novel DNA-binding drugs is described here. The method may be applied to any known binding site for any specific transcription factor, regardless of whether the identity of the transcription factor itself is known.

TATA-boxes have been determined for a large number of genes. Typically, the TATA-box consensus sequence has been identified by examining the DNA sequence 5' of the RNA start site of a selected gene. However, the most rigorous determinations of TATA boxes have also demonstrated the transcription factor binding site by DNA protection experiments and DNA:protein binding assays (using electrophoretic methods). Many of these sites are annotated in the public databases "EMBL" and "GENBANK", which both contain sequences of nucleic acids sequences. Unfortunately, the flat field listing of these databases do not consistently annotate these sites. It is possible, however, to automatically search a database, using a text parsing language called AWK, to extract most sequence information that relates to annotated promoter sequences.

The following is a description of how selected promoter sites were located in the public database from "EMBL." The flat field annotations from "EMBL" Version 32 as processed by "INTELLIGENTICS" (Mountain View, CA), were

obtained with the set of UNIX programs call "IG-SUITE." These programs were executed on a "SUN IPX" workstation. An AWK script was used to parse all the primate annotation files listed in the "EMBL" database. The AWK interpreter is supplied as part of the system software that comes with the "SUN IPX" workstation.

The following is a description of how the AWK parses annotation files looking for and printing information relating to promoters and TATA-boxes. The system is asked to examine the input files for certain key words in the header lines or annotations to the sequence. The AWK interpreter reads input files line by line and executes functions based on patterns found in each line. In this case, the AWK system read the annotation files of EMBL. The following is a description of how the AWK script can be used to parse out sequences containing TATA-boxes.

The program first examines the files for all header lines containing the word "complete" but not "mRNA" or "pseudogene"; the output is printed. Complete genes sometimes contain the promoter sequences but complete mRNA genes do not contain the promoters. mRNA genes are not of interest for the purpose of detecting promoter elements. Next, the AWK system looks for the word "exon 1" and if it finds it prints the header and "DE" line. Then it looks for "5'" and prints the header line if it does not contain the word "mRNA". Next it looks for the word "transcription" and if it finds it prints the preceding and following line along with description line.

Next, the AWK system examines the files for the word "TATA" in the header lines or references. This results is printed. After this it looks for the word "promoter" and if it finds it prints that line and the line after it which contains the information about the promoter. Then the program looks for "protein bind" and prints that line along with the next one. The description of "protein bind" is usually used to mark potential binding sites of transcription factors in the "EMBL" database. AWK then scans for any annotated primary mRNA start sites. The promoter sequence is found in front of the start site. Finally, any exon 1 start sites that are annotated in the feature table are extracted. Exon 1 start sites should also be the primary transcription start site and the TATA boxes usually are found approximately 25-35 base pairs 5' to the transcriptional start site.

The actual AWK script is included here as an example of how to parse a database to extract promoter sites:

```

BEGIN {print_next_line=0}
{if (print_next_line==1)
    {print $0
    print_next_line=0}
}
{if ($0 ~/^>/)
    { Locus=$0
    l_flag=0 }
}
/^>/ && / [Cc]omplete/ && $0 !~ /mRNA|mrna/ && $0 !~ /pseudogene/{print}
/^>/ && /exon 1[^0-9]/ {print}
/^>/ && /5'/ && $0 !~ /mRNA|mrna/ {print}
/[Tt]ranscription/ {print Locus "\n" PL "\n" $0; print_next_line=1}
{if ($0 ~/^FT/ && $0 ~ /TATA/ && $0 ~ /note/)
    {print Locus "\n" PL "\n" $0}
}
{if ($0 ~/^FT/ && $0 ~/[Tt]ranscription/ && $0 ~/\//)
    {print Locus "\n" PL "\n" $0}
}
{if($2 !~ /note/ && $2 ~ /TATA/) {print Locus "\n" $0} }
{if ($2 ~ /promoter/)
    {print_next_line=1
    if(l_flag==0)
        {print Locus "\n" $0
        l_flag=1}
    else
        print $0
    }
}
}
```



```

{if ($2 ~/protein_bind/)
  {print Locus "\n" $0
   print_next_line=1}
}
5 {if ($2 ~/prim_transcript/ && $3 !~/^1..|^<1../)
  {print Locus "\n" $0
   print_next_line=1}
}
10 {if ($0 ~/^FT/ && $0 ~/number=1[^0-9]/)
    if(PL ~/exon/){print Locus "\n" PL"\n"$0}
  }
  {PL=$0}

```

15 After the AWK script is run on the database the output is manually examined. Those sites that are clearly promoter sites are noted and nucleotide coordinates recorded. Other gene sequences are examined using the "FINDSEQ" program of "IG_SUITE" to see if the promoter sites can be determined or if the references in the database describe the promoter sequences. If so, those nucleotide coordinates are noted. At the end of this examination "FINDSEQ" is used to extract any sequences containing promoter sequences by using an indirect file of "LOCUS" names constructed using a text editor.

20 A parsing program was also written to extract each of the annotated sites from the file that "FINDSEQ" extracted from "EMBL." This program extracts the following information: the promoter site name and four numbers representing the nucleotide coordinates of where the sequence is to start, what the coordinate of the first base of the site is, the coordinate of the last base of the site and the end of the sequence to be extracted. A large batch file was constructed to automatically extract each of the promoter sites. These sequences formed the basis of Table V.

30 The Sequence Listing presents a number of sequences that are useful as test sequences in the present invention. SEQ ID NO:1 to SEQ ID NO:481 and SEQ ID NO:600 correspond to promoter targets (typically, TATA box-containing sites) for human genes. SEQ ID NO:482 to SEQ ID NO:599 correspond to promoter targets for viral genes.

Example 16

Using Normalized Values to Determine Sequence Specificity and Relative Binding Affinity

A. The Assay Mixture and Calibrator Samples.

40 The assay mixture is prepared as described in Example 10. The concentration of the components can be varied as described in the Detailed Description.

45 The assay mixtures containing both UL9 and DNA are incubated at room temperature for at least 10 minutes to allow the DNA:protein complexes to form and for the system to come to equilibrium. At time = 0, the assay is begun by adding water (control samples) or test molecule (typically at 1-5 μ M, test samples) to the assay mixtures using a 12-channel micropipettor. After incubation with drug for 5-120 minutes, samples are taken and applied to nitrocellulose filters using a 96-well dot blot apparatus (Schleicher and 50 Schuell) held at 4°C.

Calibrator samples are used to normalize the results between plates, that is, to take plate-to-plate variability into account. Calibrator samples are prepared using 2-fold serial dilutions of DNA in the assay mixture and incubating duplicate samples in one column of the 96-well assay plate. The 55 highest concentration of DNA used is the same concentration used in the screening samples. In general, calibrator samples are used in all experiments.

However, use of calibrator samples appears to be less important for experiments using blocked plates since the variability between blocked plates is lower than between unblocked plates.

60 The calibrator samples are used to normalize the values between plates as follows. The volume values (Example 10) for the calibrator samples are obtained from densitometry. Volume values are plotted against DNA concentration. The plots are examined to ensure linearity. The volume values for the points on the calibrator line are then averaged for each plate. A

factor, designated the normalization factor, is then determined for each calibrator line. When the normalization factor is multiplied by the average of the points on each calibrator line, the product is the same number for all plates. Usually, the average of the line averages is used for determining the normalization factor, although in theory, any of the line average numbers can be used. The operating assumption in this analysis is that the differences in the calibrator samples reflected the differences in adsorption for each plate. By normalizing to the calibrator samples, these variations are minimized.

Once the normalizing factor is obtained, all of the raw volume values for each of the test assays on the plate is multiplied by the normalizing factor. For example, if the following data were obtained, the process of normalization would be as follows:

TABLE XIII

PLATE NUMBER	DNA CONCENTRATION				
	0.8	0.4	0.2	0.1	Average
Plate I:	4000	2000	1000	500	1875
Plate II:	4200	2100	1050	525	1969
Plate III:	3800	1900	950	475	1781
	Average:				1875

Plate I has a normalization factor of 1; Plate II has a normalization factor of $1875/1969 = 0.95$; Plate III has a normalization factor of $1875/1781 = 1.05$. The equation used to establish these numbers is as follows: "Average average"/line average = normalization factor.

If the normalization factors are different, these factors are incorporated into the data analysis. The sample data on each plate is then multiplied by the normalization factor to obtain normalized volume values.

B. The Capture/Detection System.

A 96-well dot blot apparatus is typically used to capture the DNA:protein complexes on a nitrocellulose filter as described in Example 10.

C. Quantitation of Data.

The autoradiographs of the nitrocellulose filters are analyzed as described in Example 10.

D. Analysis of Data.

After densitometry, the data is analyzed using a spreadsheet program, such as "EXCEL." For each plate, the calibrator samples are examined and used to determine the normalization value. Then, for each test oligonucleotide, at each drug concentration and/or each time point, a normalized % score is calculated. The normalized % score (n%) can be described as follows:

$$n\% = (nT/nC) \times 100,$$

where (i) nT is the densitometry volume of the test sample multiplied by the normalization factor for the plate from which the sample was obtained, and (ii) nC is the densitometry volume of the control sample multiplied by the normalization factor for the plate from which the sample was obtained. The oligonucleotides are then ranked from 1 to 256 based on their n% scores.

While the invention has been described with reference to specific methods and embodiments, it will be appreciated that various modifications and changes may be made without departing from the invention.